Automatic Summarization of Legal Text

N. van de Luijtgaarden

First supervisor: M.P. Schraagen
Second supervisor: A.L. Lamprecht
Case company supervisor: R.W. Lucas

MSc Thesis

Universiteit Utrecht

Natural Sciences
Utrecht University
The Netherlands
August 28, 2019

Acknowledgements

I would like to express my very great appreciation to my research supervisor, for his patient guidance and constructive feedback of this paper and keeping my progress on schedule. Special thanks should be given to my supervisor at the case company, for his willingness to give his time so generously and help me with his experience in the legal field. My grateful thanks are extended to my classmates for providing me with valuable feedback and also giving me time to relax throughout my study.

I would also like to thank my family for their continous support of my work and keeping me motivated. Also, there are my roommates, who were of great support in deliberating over my problems and findings, as well as providing happy distraction to rest my mind outside of my research. In addition, I want to thank my sports association for offering me a place to relax and give me the energy to continue my passion for this research. Finally, I wish to thank the law students that extensively evaluated the results of my experiments and Pandora Intelligence for providing me with a dataset for my research.

Abstract

With the legal sector embracing digitalization, the increasing availability of information has led to a need for systems that can automatically summarize one or more documents. Current research on legal text summarization has only focused on extractive methods, which can result in awkward summaries as sentences in legal documents can be very long and detailed. In this study, we argue that due to more data being available, improved hardware and matured algorithms, the time is now right for using abstractive summarization models in the legal field.

The main goal of this thesis is to discuss how we can best apply an abstractive summarization model on a legal domain dataset. A five-phased approach was used to evaluate generated summaries based on ROUGE score, abstractiveness and through a human evaluation experiment using law graduates.

ROUGE results of our experiments are comparable to state-of-the-art studies that made use of the CNN/Daily Mail dataset. Experiments show that the model excels in rewriting the long and redundant legal sentences to much shorter ones, but does not generate many new words compared to the input document. However, the conducted human evaluation showed that not all elements needed in a summary (background, considerations, judgement) were always present together in a generated summary, and that reference summaries got better relevance scores. Still, students observed that generated summaries did contain key information about cases and preferred it to using reference summaries that only contain keywords.

Through this study, we argue that there is a lot of potential for abstractive summarization in the legal field. The quality is not on the same level as the reference summaries, but it can function as a good replacement for reference summaries that only contain keywords. For improving relevance in the generated summaries, an implementation of a network that can recognize the three core elements of a case is needed. For readability, additional post-processing in the decoding function can help recognize when sentences are cut off too early. In general, we also doubt whether ROUGE is still a good metric for evaluating abstractive summarization models, as there exists an inverse relationship between the ROUGE score and the abstractiveness of a document.

Contents

1 Introduction

In recent years, more and more documents are being digitalized. According to industry research (MarketsandMarkets, 2018), the cloud storage market size is forecasted to grow from $30.7 billion in 2017 to $88.9 billion by 2022. Advancements in data storage and cloud technology allow organizations to archive and preserve large amounts of documents while keeping costs to a minimum (Cumbley & Church, 2013; Lu, Hsieh, Chang & Yang, 2013).

Traditionally, the legal field has been an industry that is slow to adopt digital transformations. In a sector that deals with highly confidential and personal documents, one can understand that it is critical for information to remain secure. However, with the advancements in cloud technology and security, the legal sector is embracing digitalization, as it can dramatically improve time and cost savings. Still, this digitalization of documents has led to an overload in information for judges, making it seem as finding a needle in a haystack. The increasing availability of information in the legal industry has led to a need for systems that can automatically summarize one or more documents (Radev, Hovy & McKeown, 2002).

Radev et al. (2002) define a summary as a "text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". The goal of a summary is to carry over the main ideas of a document while removing redundancy.

Two key approaches exist for summarization. *Extractive summarization* involves identifying important text spans from the original source document and combining them to yield a shorter text (Hahn & Mani, 2000). This method provides a strong baseline for summarization since sentences are taken straight from the document, but lacks in freedom as it cannot create novel sequences (Paulus, Xiong & Socher, 2017). On the other hand, *abstractive summarization* refers to the process of generating new sentences that explain in more general terms what the text is about (Hahn & Mani, 2000). This method, despite being powerful in theory, proves to be difficult in practice, as abstractive algorithms require great awareness of grammar, lexicons and textual context for parsing and generation.

Several studies (Farzindar & Lapalme, 2004; Hachey & Grover, 2006; Yousfi-Monod, Farzindar & Lapalme, 2010) have experimented with text summarization in the legal field. Despite showing promising results, the works have only studied small data-sets and mostly relied on handwritten features. In addition, researchers have only studied the use of extractive summarization techniques and only few works have made use of the recent advancements in neural networks (Alschner & Skougarevskiy, 2017; Elnaggar, Gebendorfer, Glaser & Matthes, 2018). This thesis identifies two gaps in current research on text summarization. First, no research to date has made use of an abstractive summarization model in the legal context. In addition, no work on abstractive summarization has evaluated their model on documents of the size and structure of a legal document, which are considerable larger than documents used in current datasets.

The idea of using abstractive summarization in the legal field has gained a lot of

criticism in literature. In the legal context, documents tend to be long and detailed as they need to reflect all information that was needed to come to the judgement. Researchers argue that "an abstract may be less accurate and less credible because it is not a direct citation of the decision; reformulation may lead to misinterpretation of the judge's intent" (Yousfi-Monod et al., 2010). We agree with this criticism, as at the time of writing, abstractive summarization models were still in its infancy; computational power needed for these algorithms was still too low, data storage was expensive and abstractive summarization algorithms were still under-developed (Cumbley & Church, 2013; Lu, Hsieh, Chang & Yang, 2013). Also, no large datasets of legal documents were available yet.

Currently, large legal text corpora, containing over hundreds of thousands cases, have made their data available for use[1]. In addition, abstractive summarization models have matured as training has sped up in the past years by, for example, using graphical processing units (GPUs) instead of central processing units (CPUs). Also, recent results of abstractive models have shown to outperform extractive models based on ROUGE scores (Paulus et al., 2017; Chen & Bansal, 2018). Thus, this work argues that the time is right for using abstractive summarization in the legal context.

The main goal of this thesis is to discuss how we can best apply an abstractive summarization model on a legal domain dataset. Moreover, this research considers state-of-the-art models for abstractive summarization and evaluates the model using automatic evaluation as well as human evaluation measures. Also, this work introduces a dataset containing over 400,000 Dutch legal court documents, including matching summaries.

First, Section 2 discusses abstractive and extractive summarization techniques, as well as examining the current work on legal text summarization. Then, Section 3 describes the approach and experimental set-up that has been used in this research. After that, Section 4 gives an overview of the results of this study. Finally, Section 5 discusses the results and implications of this research.

2 Literature study

First, this section discusses datasets used in summarization research, as well as current methods for evaluating summaries. Then, state-of-the-art extractive and abstractive summarization models will be evaluated. Finally, studies conducted on legal text summarization will be elaborated upon.

2.1 Datasets

For machine learning algorithms to work properly, a lot of data is needed. In the case of summarization, this can be difficult to find, because datasets need to include a summary for each text as well. Table 1 gives an overview of common datasets

[1] `https://www.rechtspraak.nl/` & `https://eur-lex.europa.eu/homepage.html?locale=nl`

used for text summarization. In most cases, the model is trained on either the CNN/Daily Mail, Newsroom or Gigaword corpus. Then, it is is evaluated on the DUC-2003 or DUC-2004 dataset. All datasets, except the Gigaword and New York Times corpus, are freely available online.

Dataset	**Description**	**Size**	**Summaries**
CNN/Daily Mail	News articles	287K	Yes
Newsroom	News articles	1.3M	Yes
DUC-2003 and DUC-2004	News articles	500	Yes, 4 for each article
Gigaword	News articles	9.5M	No, only article headline
New York Times	News articles	650K	Yes

Table 1: An overview of common datasets used for summarization

The most popular dataset that has been used by studies on text summarization is the CNN/Daily Mail corpus. There are two different versions of this dataset: one uses actual entity names (non-anonymized) and the other replaces entity occurences with document-specific integer-ids beginning from 0 (anonymized) (Nallapati, Zhou, Santos, Gulcehre & Xiang, 2016). Results suggest that the non-anonymized version may result in higher ROUGE scores, as multi-word named entities lead to a higher overlap of words (Fan, Grangier & Auli, 2017; Chen & Bansal, 2018).

2.2 Evaluation

Evaluation of summaries by humans is a time-consuming process, as it involves a lot of different quality metrics such as coherence, conciseness, readability and content (Mani, 2001). Adressing these challenges, Lin (2004) introduced ROUGE, a method for automatically measuring similarity between a reference summary and a created summary, based on precision and recall. For example, having a ROUGE-1 recall score of 40% implies that 40% of the words in the reference summary are also present in the generated summary. In contrast, a ROUGE-1 precision score of 40% implies that 40% of the words in the generated summary are also present in the reference summary. Table 2 gives an overview of all ROUGE evaluation measures.

Metric	**Explanation**
ROUGE-1	Overlap of each word (unigrams)
ROUGE-2	Overlap of adjacent words (bigrams)
ROUGE-L	Overlap of longest common subsequences (longest common n-gram)
ROUGE-W	Overlap of longest common subsequences which favors consecutive matches
ROUGE-S	Overlap of any pair of words in sentence order (allowing gaps between words)

Table 2: ROUGE evaluation measures (Lin, 2004)

As both precision and recall are relevant measures to compute, it is best to always report the ROUGE score using the F_1 score. The F_1 score considers both precision and recall and is the harmonic mean of precision and recall, where the perfect value is achieved at 1 and worst at 0. This measurement can be expressed in the formula below.

$$F_1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Despite ROUGE being the standard choice for summary evaluation by studies on extractive and abstractive summarization, it suffers from major problems. For example, consider the reference sentence "Police killed the criminal" (S1) and two created subsentences: "The criminal shot police" (S2) and "The criminal police shot" (S3). In this case, sentences S2 and S3 would have the same ROUGE-1, ROUGE-2 and ROUGE-L score. However, both sentences have a completely different meaning.

Another disadvantage of using ROUGE is that perfect scores for extractive summarization are theoretically hard to achieve, as this indicates that the reference summary must introduce no new words (Schluter, 2017). In addition, supervised summarization methods tend to perform better on ROUGE scores as opposed to unsupervised approaches, because they tend to have the same writing style as the reference summary.

A common practice is summary evaluation against multiple reference summaries, and then averaging ROUGE scores across these summaries (Nenkova & Passonneau, 2004). However, when Schluter (2017) compared human summaries against each other for the DUC'2004, only a 39.92 ROUGE-1 and 9.39 ROUGE-2 was achieved. This indicates that a summary can be written in many different ways, while containing more or less the same information.

2.3 Extractive summarization

Sentence ranking is one of the earliest methods that has been used for summarization. Another popular approach is graph-based and aims to identify important sentences by comparing paths in the graphs. Finally, through neural network developments, sentence embedding has become an interesting topic for text summarization as well.

2.3.1 Sentence ranking

A common approach for sentence ranking is word frequency. Salton & Buckley (1988) and Sparck Jones (1972) suggested that the most important terms have a high word frequency in the document, but a low overall word frequency in the document corpus. Following this, term importance can be defined as the product of term fequency and inverse document frequency (tf*idf).

Carbonell & Goldstein (1998) introduced an approach for a search engine that considers both relevance to the query as well as whether it contains minimal similarity

to previously selected document. This approach of encouraging both sentence relevance and diversity was applied to summarization as well by Nomoto & Matsumoto (2001). In their work, the authors used a clustering algorithm to find diverse topic areas in the text, while identifying the most importance sentences in each of these clusters using tf*idf.

A recent study describes a quantitative and qualitative evaluation of algorithms for sentence summarization (word scoring, sentence scoring and graph scoring) found in literature (Ferreira, de Souza Cabral, Lins, Pereira e Silva, Freitas, Cavalcanti, Lima, Simske & Favaro, 2013). Results on three different datasets (news articles, scientific papers and blogs) showed that word frequency, tf x idf, sentence position and sentence length were among the top performers. Still, the authors observed that performance of algorithms was dependent on the dataset used.

Through developments in neural networks, machine learning algorithms have become more popular for sentence ranking. Cao, Wei, Dong, Li & Zhou (2015) developed a framework for multi-document summarization upon a Recurrent Neural Network (RNN) to rank sentences based on a list of word and sentence level features. In a similar study, Nallapati, Zhai & Zhou (2016) present a mechanism that allows the extractive model to be trained using human generated reference summaries alone (SummaRuNNer).

Another study presented two neural architectures based on RNNs for sentence ranking (Nallapati, Zhou & Ma, 2016). The first strategy, *Classify*, traverses to the document in the original order and decides whether each sentence belongs to the summary. The other strategy, *Select*, picks sentences one at a time, in an order that the model deems fit.

A paper by Cao et al. (2016) presented a summarization system called AttSum, which aims to generate summaries that can perform well on both query relevance ranking and sentence saliency. Their system applies an attention mechanism that simulates the reading behaviour of humans. In addition, a joint neural network model is used to learn both query relevance ranking and sentence saliency. Similar work by Zhou, Yang, Wei, Huang, Zhou & Zhao (2018) proposes a neural network to that is used to jointly learn to score and select sentences (NeuSum), while previous methods treated these as two individual tasks.

2.3.2 Graph-based

Novel work on graph-based information retrieval was conducted in a study by Page, Brin, Motwani & Winograd (1999), where the authors discuss PageRank, a method for measuring the relative importance of Web pages. PageRank works by counting the number of links on the web to the current page to determine its importance. Essentially, a link from page A to page B is considered as a vote for page B. However, when page A receives a lot of votes from other websites, their vote is deemed more important. Figure 1 shows an example of a PageRank calculation.

A similar line of thinking can be applied to text summarization to rank sentences in a document. A paper by Mihalcea & Tarau (2004) introduced TextRank, a graph-based ranking model for graphs extracted from natural language text. Following the

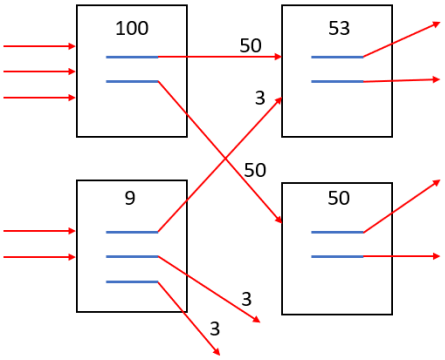

Figure 1: An example of a PageRank calculation (Page et al., 1999)

PageRank methodology, sentences that are similar to many others, are deemed as very important. The importance of this sentence is also determined by the importance of sentences that are similar to it. While TextRank creates a new similarity graph for each individual document, a study by Erkan & Radev (2004) applied the PageRank methodology to create a graph for a corpus of documents (LexRank).

2.3.3 Sentence embedding

A fundamental problem that makes machine learning difficult to apply to Natural Language Processing (NLP) is the curse of dimensionality; a sentence on which the model is tested, will likely be different from all sentences seen in training. Bengio, Ducharme, Vincent & Jauvin (2003) proposed a model that learns a distributed representation for a word which allows the model to be informed about a number of semantically similar neighbouring sentences. In later work, Mikolov, Sutskever, Chen, Corrado & Dean (2013) introduced the word2vec model based on skip-grams, which added more complex types of word relationships, such as opposites, tenses, plurals and phrases. Kiros, Zhu, Salakhutdinov, Zemel, Torralba, Urtasun & Fidler (2015) abstracted this model to the sentence level through so called *Skip-Thought Vectors*: instead of using a word to predict its surrounding context, a sentence is encoded to identify similar sentences around it. Another popular sentence embedding framework is GloVe, which leverages the statistics of the corpus to produce a vector representation for a word (Pennington, Socher & Manning, 2014). Whereas the Skip-Thought is a predictive model, GloVe is a count-based model which is based on matrix calculations. Both methods tend to perform similar for different datasets, but one benefit of using GloVe is that it is easier to train over more data. In more recent work, Devlin, Chang, Lee & Toutanova (2018) created BERT (Bidirectional Encoder Representations from Transformer), a contextual language model that generates a representation of each word, based on other words in the sentence. Context-free models such as GloVe and Skip-Thought Vectors generate a single word embedding for each word, which means that words such as bank have the same representation, even when found in sentences containing *bank deposit* or *river bank*.

Sentence embeddings can be used for text summarization in an unsupervised manner. One approach is to cluster sentences with similar topics and select the most

central sentences from each cluster for your summary (Padmakumar & Saran, 2016). An advantage of using this method is that it guarantees that different topics from the text are discussed in the summary. Cheng & Lapata (2016) proposed a model that uses a RNN to extract important sentences by relying on sentence embedding techniques (NeuralSum). Table 3 provides an overview of the extractive summarization models that have been discussed in this section and their results.

Technique	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Dataset**
TextRank (Mihalcea & Tarau, 2004)	49.04%	-	-	DUC-2002
LexRank (Erkan & Radev, 2004)	37.36%	-	-	DUC'2004
AttSum (Cao et al., 2016)	43.92%	11.55%	-	DUC'2007
Sentence embedding(Padmakumar & Saran, 2016)	35.74%	6.71%	-	Tipster
NeuralSum (Cheng & Lapata, 2016)	42.20%	17.30%	34.80%	CNN/Daily Mail
SummaRuNNer (Nallapati et al., 2016)	42.00%	16.90%	34.10%	CNN/Daily Mail
Classify/Select (Nallapati et al., 2016)	42.20%	16.80%	35.00%	CNN/Daily Mail

Table 3: An overview of state-of-the-art extractive summarization methods. Results are reported in the ROUGE recall score, as most extractive summarization studies have not reported the ROUGE F_1 score. All studies that trained and evaluated using the CNN/Daily Mail dataset made use of the anonymized version

2.4 Abstractive summarization

As discussed before, abstractive summarization is defined as the process of generating novel sentences that explain in more general terms what the text is about (Hahn & Mani, 2000). Instead of identifying important sentences and copying them, abstractive techniques create new sentences and thus have more freedom in their summary and are closer to what a human might write. However, these techniques are difficult to implement in practice and require the use of complex neural networks.

The first study on abstractive summarization combines a neural language model with an attention-based input encoder which makes no prior assumptions on the document corpus (Rush, Chopra & Weston, 2015). In a later study, Chopra, Auli & Rush (2016) improved on this summarization model by changing the attentive recurrent architecture.

Nallapati et al. (2016) proposed a pointer network that can be used to properly strike a balance between being faithful to the original source (named entities) and allowing creativity. In addition, applying linguistic features such as tf x idf help to identify key concepts and entities. Li, Lam, Bing & Wang (2017) managed to improve slightly on this model by implementing a latent structure modeling component in the recurrent

neural network, organizing the text in structures such as “Who”, “What Happened” and ”Why”.

One shortcoming of abstractive summarization methods was that they often reproduce factual details inaccurately. To solve this, See, Liu & Manning (2017) implemented a hybrid pointer-generator network that can copy words from the text through pointers, but also has the ability to create novel words through the generator. Still, one problem that remained for abstractive summarization was that for longer documents and summaries, models tend to include repetitive and incoherent phrases. A study by Paulus et al. (2017) implemented an intra-attention mechanism for multi-sentence summarization that evaluates input and output continuously, in order to look at words that have already been generared by the decoder. Also, the authors added an unsupervised learning method to the neural network to improve readability of the summaries.

A recent study by Chen & Bansal (2018) selects important sentences and then rewrites them abstractively (compresses and paraphrases) in order to generate a summary. Gehrmann, Deng & Rush (2018) explores the use of a bottom-up attention step to make abstractive summarization more efficient. Their approach is able to effectively compress sentences, while still generating fluent text. Al-Sabahi, Zuping & Kang (2018) have implemented a bidirectional RNN that enables the model to handle both history and future textual context to generate multi-sentence summaries. Table 4 gives an overview of all discussed abstractive summarization methods and what results they have achieved so far.

Technique	ROUGE-1	ROUGE-2	ROUGE-L
Bidirectional Model* (Al-Sabahi et al., 2018)	**42.60%**	**18.80%**	38.50%
Fast Abstractive Rewriting** (Chen & Bansal, 2018)	40.88%	17.80%	**38.54%**
Bottom-Up Model (Gehrmann et al., 2018)	41.22%	18.68%	38.34%
Facebook: Controllable Abstractive Summarization** (Fan et al., 2017)	40.38%	17.44%	37.15%
Google: Pointer-Generator** (See et al., 2017)	39.53%	17.28%	36.38%
Fast Abstractive Rewriting** (Chen & Bansal, 2018)	39.66%	**15.85%**	**37.34%**
SalesForce: Reinforced Learning** (Paulus et al., 2017)	**39.87%**	15.82%	36.90%
Facebook: Controllable Abstractive Summarization** (Fan et al., 2017)	39.06%	15.38%	35.77%

Table 4: An overview of state-of-the-art abstractive summarization methods. Results are reported in the ROUGE F_1 score. The dataset used is the CNN/Daily Mail corpus. Studies in the upper part of the table make use of the non-anoymized dataset, while the studies in the below part of the table use the anonymized dataset. (*) = code is open-source, (**) = code is open-source and human evaluation is included. Best results achieved for each metric and dataset are shown in bold

2.5 Legal text summarization

It is only since the work of Farzindar & Lapalme (2004) that the study on legal text summarization has gained momentum. Here, the authors proposed LetSum, a system that uses thematic structures and the document's architecture to create a table style summary. Their approach mostly relied on hand-written dictionaries and heuristic features such as the inverse document frequency.

Moens & de Busser (2002) experimented with a strategy for the summarization of scientific articles by adding the rhetorical status to each sentence in a document. Using this approach, Grover, Hachey & Korycinski (2003) describe a method to identify the rhetorical structure of sentences in documents of civil cases from all of the United Kingdom and criminal cases from England, Wales and Northern Ireland. In a later study, these rhetorical structures were used to determine the best sentences to include in a sumary using several machine learning algorithms (Hachey & Grover, 2006). This was one of the first studies that summarized legal text without the use of hand-crafted dictionaries.

Saravanan, Ravindran & Raman (2006) discussed an approach for applying probablistic graphical models based on Conditional Random Fields (CRF) for automatic text summarization. In a later study, Saravanan & Ravindran (2010) applied their approach to different sub-domains of court documents, such as rent control, income tax and sales tax. The authors observed that performance of the summarization system did not change across the different sub-domains.

Yousfi-Monod, Farzindar & Lapalme (2010) used a Naive Bayes algorithm with a set of heuristic features to identify sections (introduction, context, reasoning, conclusion) and create a summary. They found out that the quality of the summary differed for individual sections (introduction, facts etc.), indicating that for some sections it is more difficult to identify important text spans.

Technique	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Dataset**
Farzindar & Lapalme (2004)	58.00%	-	-	3.5K Canadian Law cases
Saravanan et al. (2006)	80.00%*	-	-	200 Indian Law documents
Hachey & Grover (2006)	25.50%*	-	-	200 UK Law documents
Yousfi-Monod et al. (2010)	-	64.70%*	-	4K Canadian Law cases
Galgani et al. (2012)	29.10%	-	-	3K Australian Federal Court cases
Merchant & Pande (2018)	58.00 %	15.00%	35.00%	50 Indian Court documents
Elnaggar et al. (2018)	82.00%	75.00%	82.00%	20K European Parliament legislative documents

Table 5: An overview of studies on legal text summarization. Results are reported in the ROUGE recall score; only scores marked with a (*) used the ROUGE F_1 score.

A paper by Galgani, Compton & Hoffmann (2012) discusses an approach based on the Ripple Down Rules (RDR) methodology (Compton & Jansen, 1990), where rules to identify relevant text (catchphrases) are created by a domain expert without the involvement of a knowledge engineer. Twenty three rules were defined, based on heuristics such as term frequency, citations to other cases, sentence length and Part-Of-Speech (POS) tags.

Created catchphrases were individually compared with reference catchphrases. When both catchphrases' similarity was above a certain threshold (for example, 10/15 words), then it was considered a match. By dividing all matched catchphrases by the total amount of catchphrases, the ROUGE-1 recall value could calculated. The precision value represented the number of catchphrases divided by the total number of sentences that were extracted in the first place.

Two studies by Sulea, Zampieri, Malmasi, Vela, Dinu & van Genabith (2017) and Sulea, Zampieri, Vela & van Genabith (2017) explored the use of text classification for predicting the decisions of French Supreme Court cases. The authors used a

dataset of around 130K cases and made use of a combination of hand-crafted features and linear classifiers. Conrad & Al-Kofahi (2017) also made use of a very large dataset for text extraction, as they used a text corpus containing approximately 400K documents of jury verdicts and settlement records.

In a later study, Chalkidis, Androutsopoulos & Michos (2017) extracted contract elements (title, parties, dates etc.) using several linear classifiers algorithms. A set of 3.5K English contracts that were labeled with annotations was used for training, while a larger set of 750K unlabeled contracts were used to create word embeddings using word2vec. Also, hand-crafted features and POS tags have been used to extract contract elements. The authors achieved an F_1 score of 80%, when comparing individual tokens with the labeled dataset. To date, this is the largest dataset that is used for text classification in the legal area.

It was not until the work of Alschner & Skougarevskiy (2017) that researchers made use of neural networks for NLP tasks on legal text. The authors relied on a corpus of 23K investment treaties to train a RNN to generate drafts between states based on prior treaty practices. In a small case study, they tried to predict the outcome of long-standing negotiations between China and the United States. By simulating the bargaining power of each country, the model is able to find a consensus for both countries, based on their history of investment treaty practices.

Most of the previous work only focused on text extraction in the legal area. Recent research by Merchant & Pande (2018) proposed an automated text summarization system that makes use of latent semantic analysis (LSA) to capture concepts in a legal document. LSA is an unsupervised technique that is similar to sentence embedding that was discussed in Section 2.3.3.

Elnaggar, Gebendorfer, Glaser & Matthes (2018) used the MultiModel algorithm of Kaiser, Gomez, Shazeer, Vaswani, Parmar, Jones & Uszkoreit (2017) for translation, summarization and classification through transfer learning. For summarization, the authors made use of a dataset containing around 20K legislative documents of the Europen Parliament since 1958. Each document is labeled with a short description (1-3 sentences represting the core). Results showed that when first trained on another task (translation or classification), the model also performed better on summarization, as opposed to starting from scratch.

3 Methods

This section describes the approach that has been used in this research. First, this paper discusses the goal of our summaries and the summarization model chosen for this research. After that, the used dataset will be described and the architecture and deep learning pipeline of this study will be considered. Also, the evaluation methods and set-up for our experiments examined.

3.1 Goal of our summary

Before choosing an abstractive summarization model, it is important to define the stakeholders for this research. This determines the goal of our summary, and thus the main points that it needs to bring over. We define our stakeholders as legal experts at a consulting company that give advice on legal related issues. For these stakeholders, a summary can be useful for quickly finding relevant cases and their final judgements. More specifically, this group does not desire as much detail as lawyers, but do require specific information of the case to give a proper analysis to clients. Still, it is very important for the model to retain core facts found in sentences, especially in such a critical area as the legal field. In preliminary discussions with the stakeholders, we have defined three topics that need to be included in the summary. First, the main question or problem of the case should be discussed. Then, the most important facts and considerations of the case need to be included. Finally, the judgement of the case has to be considered. Ideally, a summary should be around six sentences long.

3.2 Choosing a model

As discussed in Section 2.4, many different models exist that can be used for summarization. This section will evaluate some of these models (Paulus et al. (2017); Fan et al. (2017); Chen & Bansal (2018)) more in-depth, in order to choose the best approach for this research. The reason for choosing these three studies is that they are all open-source, conducted a human evaluation as well as they are produced in the last few years. Having an open-source codebase is important, because this allows us to make changes in the code, so we can make adjustments for legal cases, as they are very different from news articles. In addition, by making improvements to the code, we can also contribute to the general field of abstractive summarization.

Human evaluation is also very important, as a careful and thorough evaluation of summaries is needed in such a critical sector as the legal industry. Finally, these models have achieved one of the highest ROUGE-2 and ROUGE-L values on the CNN/Daily Mail dataset. Figure 2 shows a high-level overview of criteria used for evaluating an abstractive summarization model.

Appendix D shows the results of our evaluation and Appendix A discusses the algorithms of the chosen models in more detail. Based on performance, the model of Chen & Bansal (2018) stands out. However, the model of Paulus, Xiong & Socher (2017) uses the non-anonymized CNN/Daily Mail dataet, which has shown to lead to lower ROUGE scores of about 1-2%. Training time is unfortunately not known for two of the models, but Chen & Bansal (2018) claim that they are faster when compared to some of the more older models, such as the Pointer-Generator network by See et al. (2017).

Human evaluation was conducted based on readability and relevance by Chen & Bansal (2018) and Paulus et al. (2017), while Fan et al. (2017) only look at preference. Here, it is difficult to compare the models, as all methods were slightly different. All authors compared their model to the model of See, Liu & Manning (2017), one of the older models in abstractive summarization.

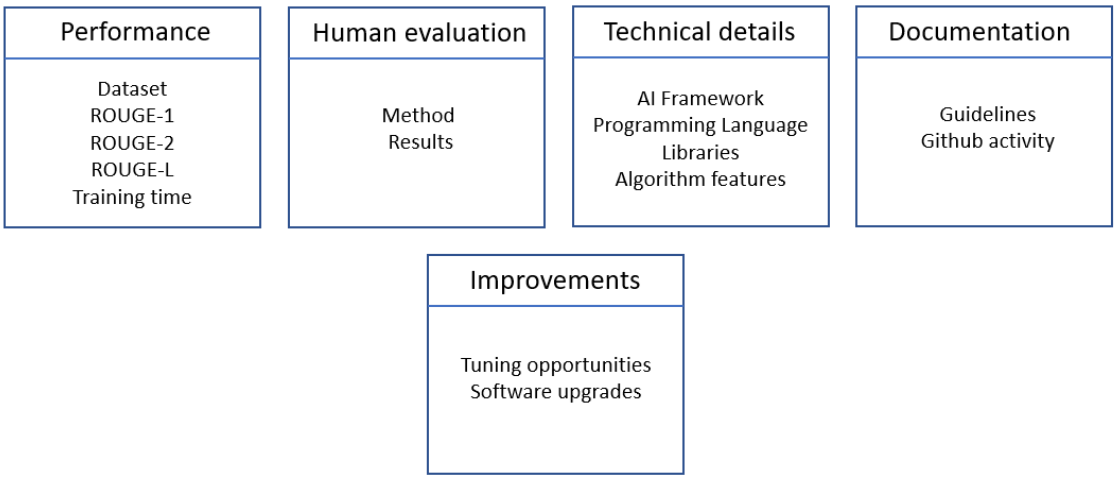

Figure 2: High-level overview of criteria

Both Chen & Bansal (2018) and Fan et al. (2017) make use of the PyTorch framework, while Paulus et al. (2017) use Tensorflow. There has been a lot of discussion in the community around which one is best, but it seems that both are equivalent in terms of networks and architecture. Still, Tensorflow has been making a lot of progress in the past year, especially with the release of Tensorflow 2.0 coming up. The model of Chen & Bansal (2018) implements the reinforcement learning framework of Paulus et al. (2017) in their approach, which has led to several good improvements. On the other hand, Fan et al. (2017) have a novel approach where many options are available to determine summary length and what entities the model needs to focus on.

By far, the documentation of the model of Chen & Bansal (2018) is most comprehensive. In addition, one of the authors is still very active on their Github repository. Fan et al. (2017) have little to none documentation about the actual implementation of their model. Paulus et al. (2017) have not documented and open-sourced their code themselves, which means that the implementation of the other authors (Keneshloo, Shi, Ramakrishnan & Reddy (2018)) might be slightly different. In addition, not much documentation is currently available here.

For the model of Paulus et al. (2017), one of the potential improvements is to rewrite the code to a newer version of Tensorflow and make better use of our GPU by using CUDA 10. The same counts for the model of Chen & Bansal (2018) and Fan et al. (2017), where we could make use of the newest Torch version and CUDA 10, which have shown to increase performance. For Chen, the vocabulary is still small, so we could try to expand this, as other studies use a much larger vocabulary (up to 150K tokens).

The core value of the algorithm by Chen & Bansal (2018) is interesting for our research, as it uses both an extractive and abstractive model. This helps us to retain the core facts found in the legal field, while also making sure the text is shortened and more readable. Paulus et al. (2017) provides an interesting approach where the model look sback at previously generated inputs and makes sure the model does not repeat itself. However, Chen & Bansal (2018) have already implemented this technique in their model. Fan et al. (2017) provide option to determine the desired

length of the summaries and what entities the model needs to focus on. Still, the application of these options proves to be limited in practice.

In this research, we choose to implement the model of Chen & Bansal (2018). The low training time makes it easy for us to tune the model and handle the large documents (up to three times larger than news articles) found in the legal field. In addition, despite having the highest ROUGE score, we can still see there are many opportunities to improve this. The core value of the algorithm helps us to keep important facts of the case in our summaries. Finally, the documentation is comprehensive, which helps us to tune the model to the needs of the study, as the model is currently optimized for the CNN/Daily Mail dataset.

3.3 Data exploration

On average, around 1.6M cases are managed by the Dutch judicial system every year. Out of these cases, a small percentage is published on Rechtspraak, a free-to-use website that can be used to find and read legal judgements. One advantage of using Rechtspraak is that they offer a free-to-use API to retrieve cases, as well as an option to download an archive of more historic cases. For this research, a pre-processed version of the Rechtspraak data provided by Pandora Intelligence[2] is used, consisting of a file that contains the type, summary and verdict of each case. In total, this file contains close to 430K legal court cases. However, only 94% of these cases contain a summary, and we only included these cases in the data exploration of this section.

Property	**Rechtspraak**	**CNN/DailyMail**
Number of documents	403,585	311,672
Words	945,008,560	238,740,752
Sentences	56,760,871	9,269,125
Average number of words	2341.54	766
Average number of sentences	140.64	29.74
Average number of words in a sentence	16.65	25.75
Average number of words in a summary	62.07	53
Average number of sentences in a summary	3.41	3.72

Table 6: Metadata for the CNN/Daily Mail and Rechtspraak dataset

Table 6 gives an analysis for both the CNN/Daily Mail and Rechtspraak dataset. One thing to notice is that the Rechtspraak dataset is much larger, as it contains more documents and words. For example, a document from Rechtspraak contains almost three times as much words as a news article. On the other hand, the average number of words in a sentence on Rechtspraak is much shorter. However, we argue that this occured because legal documents contain a lot of enumerations and headers.

[2] `https://www.pandoraintelligence.com/`

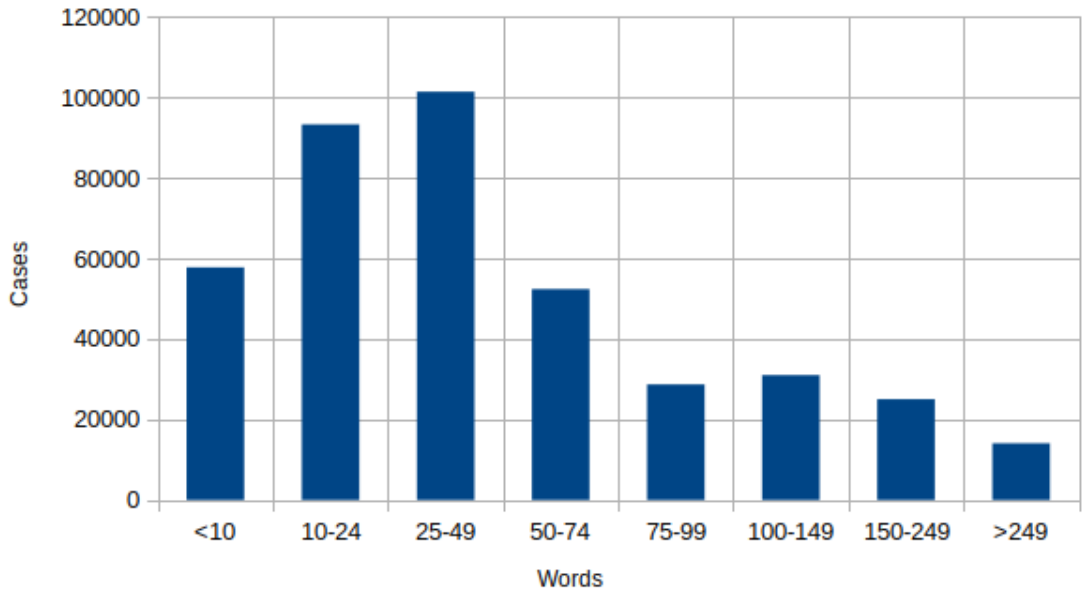

Figure 3: Word distribution for summaries of Rechtspraak cases

Looking at Table 6 we can see that the average number of sentences in a summary on Rechtspraak is almost the same as for the CNN dataset. However, as illustrated in Figure 3, close to 150K cases on Rechtspraak have less than 25 words; these cases only contain keywords or a single sentence. Our dataset mostly contains cases from the last decade, as shown in Figure 4. We consider this as a good thing, as language and style of documents can change a lot when, for example, comparing recent documents with documents of thirty years ago.

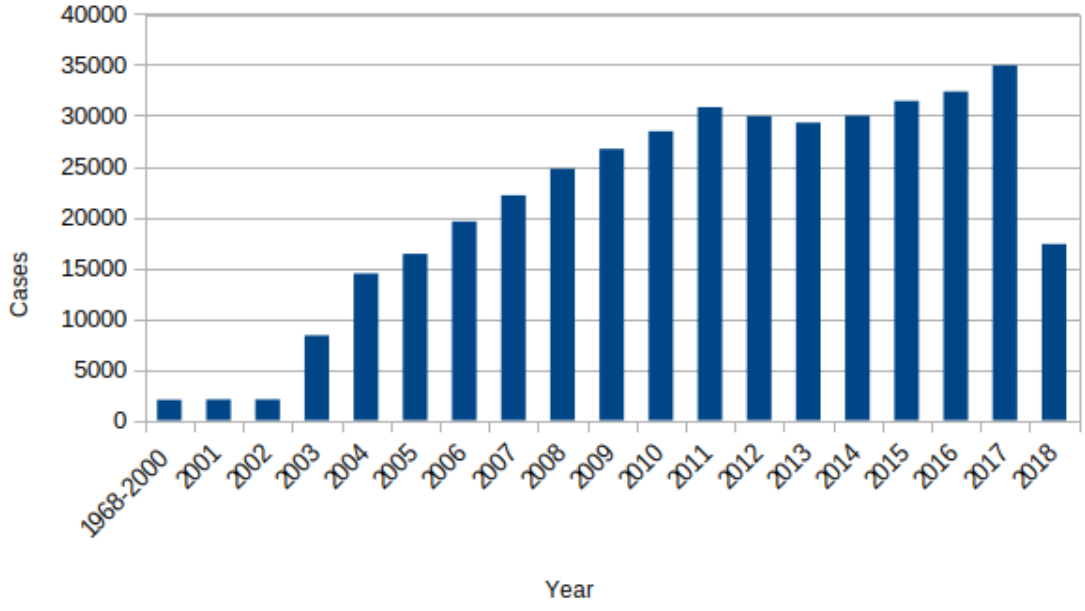

Figure 4: Year distribution for Rechtspraak cases

Table 5 shows the distribution of case categories. The tree main categories are

Civil Rights, Administrative Law and Criminal Law, while others are more specific subcategories. Thus, a case can be part of several different subcategories. However, some cases can be part of different main case categories as well. Here, you can see that Administrative Law and Civil Rights are by far the biggest categories in the legal corpus.

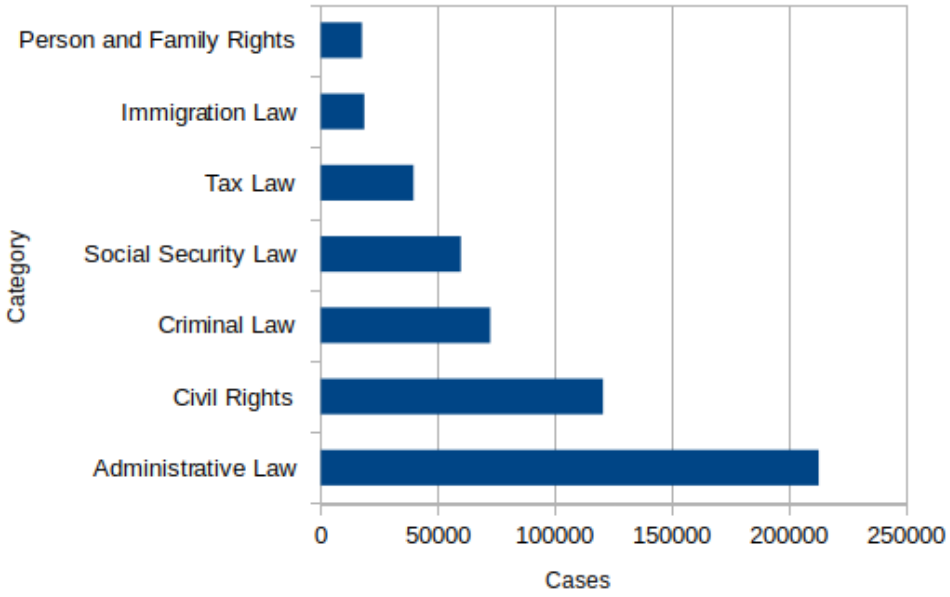

Figure 5: Category distribution for Rechtspraak cases

Figure 6 shows the distribution of words for each of the categories. Here we can see that cases of the Civil Rights category are generally much shorter compared to other categories. However, for the most part, the distribution is equal to the general word distribution found in Figure 3.

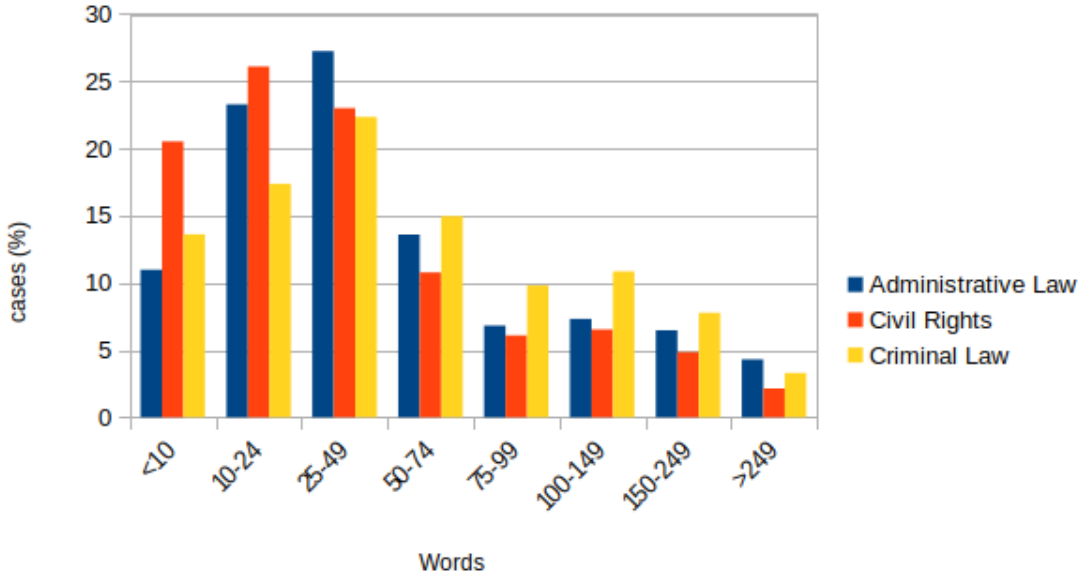

Figure 6: Word distribution per category for summaries of Rechtspraak cases

3.4 Deep learning pipeline

Figure 7 shows the pipeline for this research based on the model of Chen & Bansal (2018). In the first step, data from Rechtspraak is loaded based on a set of filters (minimal number of words, type of case etc.) and tokenized using *Ucto* (van Gompel, van der Sloot & van den Bosch, 2012) and *Stanford CoreNLP* (Manning, Surdeanu, Bauer, Finkel, Bethard & McClosky, 2014). These two frameworks ensure that the text in the verdict and summary are split properly in separate sentences. *Gensim* is used to create word embeddings through Word2Vec, while *pandas* is used to manage cases in our data file.

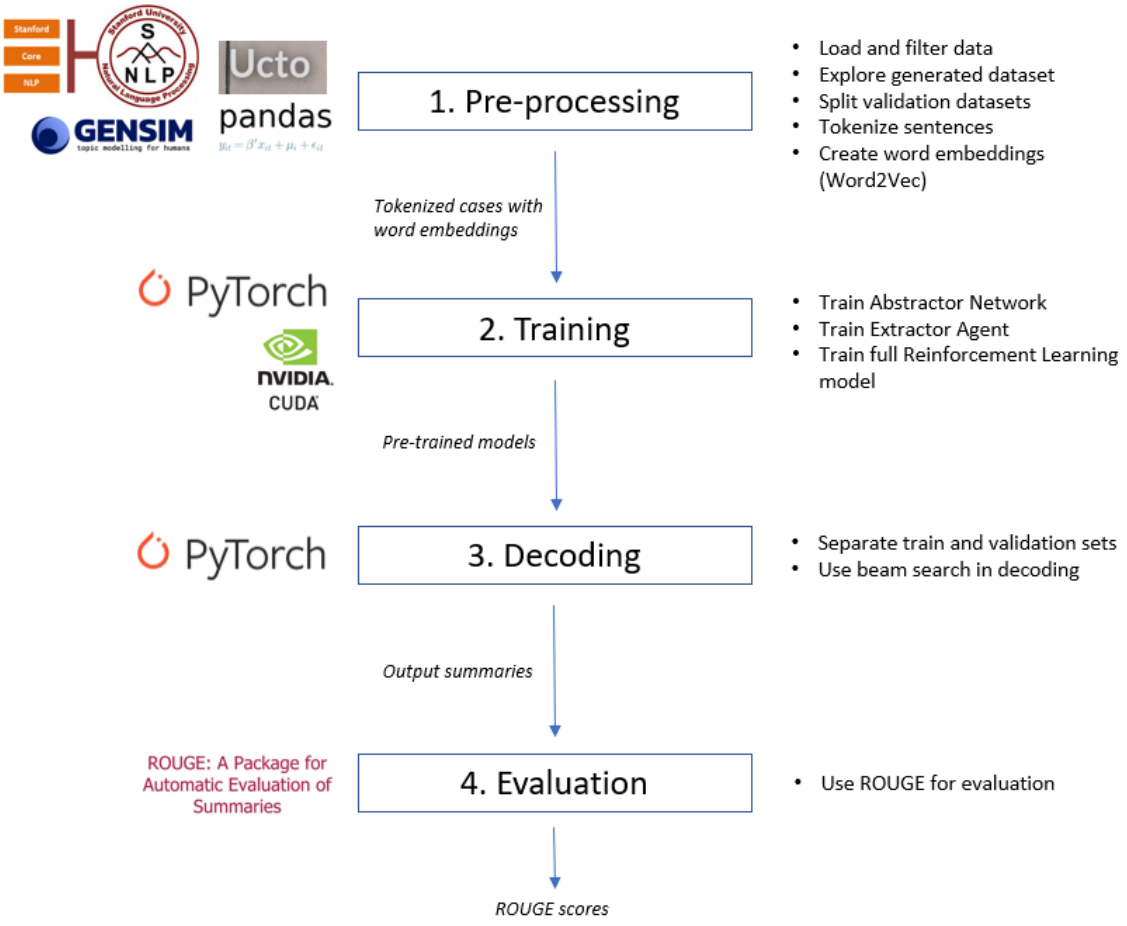

Figure 7: Pipeline

In the second step our model is used to train an abstractor network and extractor agent based on the PyTorch framework and CUDA, a parallel computing platform created by Nvidia. A full reinforcement learning model, using techniques from Paulus et al. (2017), can then be trained using the abstractor network and extractor agent. Using the trained models from this step, new summaries can be generated in step three. In step four, summaries are evaluated using the ROUGE metric.

3.5 Evaluation

Table 7 gives an overview of the evaluation phases in this research. For the first phase, the model needs to be upgraded to use the latest versions of Python, Pytorch, Nvidia CUDA and other software packages. The artificial intelligence landscape progresses fast, so many improvements are made to the architecture of the framework since the release of the paper by Chen & Bansal (2018). Also, the tokenizer needs to be changed to make it more specific to the Dutch language, as the CNN/Daily Mail dataset is in English.

Phase	**Task**
1. Initial model	Set-up the model and perform a manual evaluation of initial results
2. Improved model	Improve tokenization and pre-processing of cases. In addition, change filters in the datasets such as number of words and sentences. Goal is to improve ROUGE scores of phase 1
3. Specific evaluation	Evaluate on different categories of cases and year ranges in your test set and compare ROUGE scores
4. Abstractiveness and technical evaluation	Calculate the abstractiveness (novel n-gram count) of our model compared to the input document and reference summary. Also, compare ROUGE scores, training time and abstractiveness when changing the number of cases
5. Human evaluation	Asses readability and relevance of generated summaries and compare with reference summaries

Table 7: Evaluation phases

In phase 2 the goal is to improve tokenization and pre-processing of cases based on a manual evaluation of the results of the previous phase. We also look into changing filters in the dataset, such as the number of words and sentences. Goal is to improve ROUGE scores, but solving problems found in the manual evaluation of the initial model have a higher priority to be improved. In this phase we also conduct a small qualitative evaluation to assess the behaviour of our model.

In the next phase we use the model from the previous phase to evaluate on types of cases (e.g. Administrative Law) and on other year ranges (e.g. cases before and after 2000). In the fourth phase, an evaluation of the abstractiveness of our model is conducted. Also, we look at the ROUGE score, training time and abstractiveness when changing the number of cases. In this way, we can see how a larger training set affects the two main automatic measurements of our experiments: ROUGE and abstractiveness. Here, the same hyperparameters are used as in phase 2.

3.5.1 Human evaluation

In the final phase, a human evaluation is conducted to assess the readability and relevance of our generated summaries. Relevance is based on whether the summary contains all important information of the input article and whether it avoids generating repeated and redundant information. On the other hand, readability assesses

the fluency, grammaticality and the length of the summary. To evaluate both these criteria, we have designed an experiment: we randomly select samples from our test set and have law students rank between summaries (relevance and readability) generated by our model and the reference summaries of Rechtspraak. The student does not know which summary is generated by our model. This experiment followed a similar approach to previous work on evaluating abstractive summarization models (See et al., 2017; Chen & Bansal, 2018). In this experiment, two law students were assigned to perform an evaluation lasting a maximum of three hours. The evaluation experiment is divided into two stages.

In the first stage, the student receives five different cases from our test set. First, the student gets fifteen minutes to read and study a case. Then, the student is given both the generated summary and the reference summary from Rechtspraak. Now, the student rates both summaries on relevance (1-10) and readability (1-10) according to the definitions we have defined at the start of this section. In addition, the student gives a short explanation for his answers.

In the second stage, the student also receives five different cases from our test set and gets twenty minutes to read and study each case. Now, the student is given both the generated summary and a reference summary from Rechtspraak which only contains a set of keywords. Now, the student needs to evaluate which of the two summaries he would prefer to use if he were a legal expert at a consulting company that give advice on legal related issues, as descibed in Section3.1. In addition, the student evaluates our generated summary on relevance (1-10) and readability (1-10) and gives a short explanation for his answers. The results on readability and relevance from the second stage will only be used in our qualitity evaluation.

The idea behind this approach is that we can see what the quality of our summary, related to the stakeholders defined in Section 3.1. In addition, no previous work has evaluated the reference summaries of Rechtspraak before. As illustrated in Figure 3, close to 37% of all reference summaries contain less than 25 words; only a few keywords or one sentence. Using the results of the second stage, we can evaluate if our generated summaries are of better quality than the summary of these cases. Appendix F.2 and F.3 respectively contain the cases and questions documents that were used in this experiment. Due to the size of the cases, we have only included one out of five cases for both our evaluation stages.

3.6 Experimental set-up

Experiments are conducted on a machine running Ubuntu 18.04.2 LTS 64-bit[3]. The model in this project is re-written to use Python 3.7.3 and PyTorch 1.1 with CUDA 10.2 enabled, upgrading it from Python 3.6 with PyTorch 0.4.0 and CUDA 8. We make use of both *Ucto 0.17* (van Gompel et al., 2012) and *Stanford CoreNLP 3.9.2* (Manning et al., 2014) for tokenization and sentence splitting.

Results are evaluated using standard ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004) on full length F_1 following previous studies (Nallapati et al., 2016; See et al.,

[3]An Amazon Web Services (AWS) instance with an Nvidia Tesla V100 GPU was used. More details on our AWS architecture can be found in Appendix C.

2017; Paulus et al., 2017). For evaluation purposes, we also use both ROUGE recall and precision in our experiments. We divide our dataset into separate training, validation and test sets. More specically, we use a random split of 70% (training), 15% (validation) and 15% (test) cases. Hyperparameter tuning is performed on the validation set, while all experiments use the test dataset.

3.6.1 Hyperparameter details

Table 8 gives an overview of the hyperparameters used in this research. For training, we used batches of 4 samples and set the checkpoint frequency (number of update steps for checkpoint and validation) on 3000 for the abstractor/extractor network and 300 for RL training. While Chen & Bansal (2018) uses a checkpoint frequency of 1000 and a batch size of 32 for RL, we lower these values to speed up training time and save more intermediate results while training. In addition, documents from Rechtspraak have three times as much words as documents in the CNN/Daily Mail dataset, which makes a lower batch size appropriate. A lower batch size did not work for the abstractor and extractor network due to the structure of the convolutional network. Because of this, we followed the approach of Chen & Bansal (2018) and used a batch size of 32.

Parameter	**Value**
Layers	2
Hidden units	256
Batch size (abstractor & extractor)	32
Batch size (RL)	4
Checkpoint frequency (abstractor & extractor)	3000
Checkpoint frequency RL	300
Adam optimizer ML learning rate	0.001
Adam optimizer RL learning rate	0.0001
RL discount factor	0.95
Gradient clipping	2.0 (2-norm)
Word2Vec dimensions	128
Word2vec vocabulary	30000
Sentence generatrion limit	30
Beam search size	5

Table 8: An overview of hyperparameters used

Learning rate for the Adam optimizer is set on 0.001 for maximumlikelihood (ML) objectives and 0.0001 for RL training. We set the discount factor for RL on 0.95 and halve the learning rate when validation loss stops decreasing, in order to speed up convergence (Chen & Bansal, 2018). Gradient clipping is used to prevent exploding gradients and uses a 2-norm of 2.0 and for all LSTMs. We use a network of 256 hidden units with one layer.

A word2vec model of 128 dimensions and 30K vocabulary (most common words) is trained on the same dataset and is updated during training. The sentence generation limit is set on a maximum of 30 tokens. When generating summaries, beam search

size is set to the maximum size of 5. A lower beam search would result in faster training times, but would lead to less optimal (greedy) results.

4 Results

First, this section discusses the results of the initial model, as well as what changes were made to improve this model. Then, an evaluation based on cases with a specific category or within a certain year range will be given. After this, the abstractiveness of our model and how the ROUGE score changes over the number of cases will we described. Finally, a human evaluation is given.

4.1 Initial model

For the initial version of the model, we filtered the dataset to use cases that have a summary containing between 40 and 150 words and 3 to 6 sentences. In addition, a minimum of 7 words was used in each sentence, because this could help to learn the model to better recognize the contexts of words. Tokenization and sentence splitting was done using *NLTK* in this phase. Finally, we ended up with a dataset of 35118 cases. Total training time was 3.5 hours (extractor: 1.5, abstractor: 1 and RL: 1).

Type	**Recall**	**Precision**	F_1 **score**
ROUGE-1	36.25 (35.90, 36.60)	44.80 (44.40, 45.20)	38.00 (37.72, 38.30)
ROUGE-2	15.17 (14.84, 15.48)	19.30 (18.90, 19.70)	16.10 (15.77, 16.44)
ROUGE-L	33.00 (32.68, 33.35)	40.88 (40.48, 41.28)	34.63 (34.35, 34.91)

Table 9: ROUGE scores for the initial model based on a 95% confidence interval

The ROUGE scores for this phase are described in Table 9. Compared to the results of the model of Chen & Bansal (2018) on the CNN/Daily Mail dataset, a higher score is only achieved for ROUGE-2 (F_1 score). Another observation is that the precision score is much higher than the recall score, indicating that the model tends to be careful with adding new words to the summary that may not found in the reference summary. The low recall value tells us that there are quite some words found in the reference summary, but not discussed in the summary. Figure 9 displays the word distribution of the generated summaries. Our model created an average of 4 sentences and 71 words for each summary.

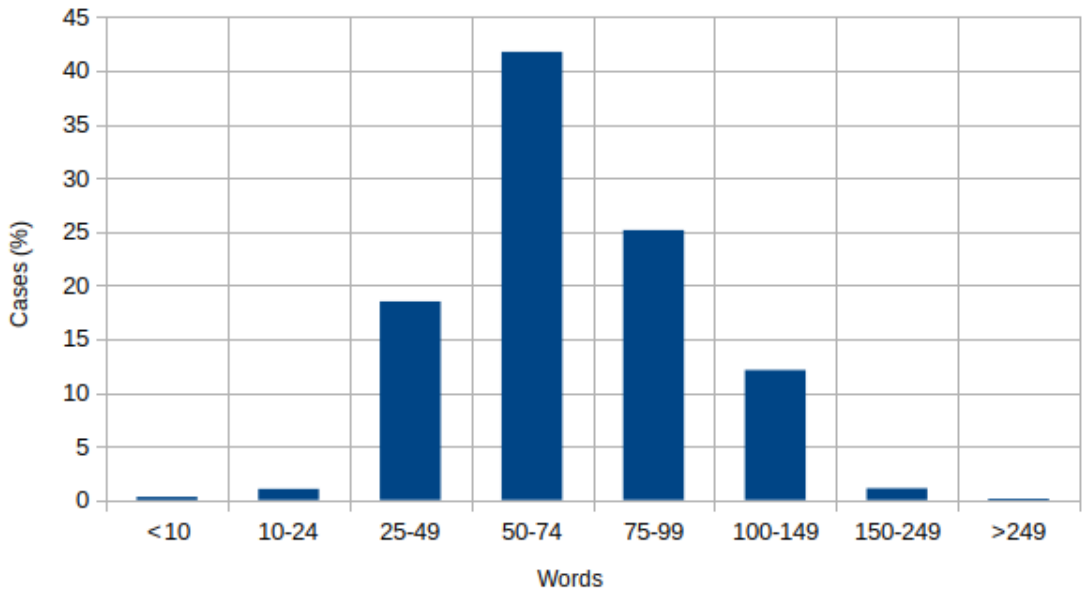

Figure 8: Word distribution for generated summaries

After performing a manual evaluation of the generated summaries, there were some patterns we noticed. First, most summaries included the key points of the verdicts and did not go in too much detail. Also, the model shortened some of the long sentences found in cases, removing redundant information while preserving the core issues. Still, sentences did not flow well with each other and the grammar was not perfect. The model also sometimes shortened sentences too early, resulting in sentences that do not have much meaning. Finally, we notices that Dutch abbreviations such as 't.o.v', 'dd.' or 'c.q.' were not handled properly by our tokenizer, resulting in defects in our training set.

4.2 Improved model

In this phase we aimed to improve the ROUGE results by changing the way the data is pre-processed. Also, the goal is to challenge some of the issues found in the manual evaluation of the initial model. As discussed in the previous section, one of the errors made by our tokenizer was that abbrevations are not treated properly. This lead to sentences not being split properly, and thus our model also created sentences that would suddenly stop. Because of this, we rewrote our tokenizer to use *Ucto* instead of *NLTK*. More specifically, we use of the 'nld-historical' configuration for *Ucto* developed by Nederlab[4]. Originally, this is a configuration file for historical text. We make use of this configuration because it is more inclined to keep certain punctuation attached to words, which happens often documents in the legal field.

Just as in the previous phase, we include cases that have a summary containing between 40 and 150 words, and have between 3 and 6 sentences. However, we changed the minimal words needed in a sentence to 5, to include more cases in our training set. After filtering and pre-processing, we ended up with a dataset of 47689

[4] `nederlab.nl`

cases. Loading and filtering the model took close to three hours, which means *Ucto* is far less efficient compared to *NLTK*, which only took an hour. Total training time was 5 hours (extractor: 1.5, abstractor: 1 and RL: 2.5).

Type	**Recall**	**Precision**	**F_1 score**
ROUGE-1	43.87 (43.42, 44.29)	37.13 (36.74, 37.51)	37.24 (36.91, 37.54)
ROUGE-2	19.30 (18.85, 19.77)	16.11 (15.71, 16.49)	16.20 (15.83, 16.55)
ROUGE-L	40.23 (39.79, 40.67)	33.94 (33.55, 34.32)	34.07 (33.75, 34.40)

Table 10: ROUGE scores for the improved model based on a 95% confidence interval

Table 10 shows the ROUGE scores for this phase. On both ROUGE-1 and ROUGE-L, the initial model scores better when looking at F_1 scores. Another observation is that the recall values are much higher and precision values much lower. The behaviour of our model has changed and is, essentially, more greedy. More words are being generated and the model is less careful with adding words that may not be in the summary, compared to the initial model.

This argument can be supported by Figure 8, as our model created an average of 6 sentences for each summary and an average of 108 words. Respectivally, this is a 50% and 46% increase in sentences and words compared to the previous phase. We argue that longer summaries are generated because of the improved sentence splitting due to the usage of the *Ucto* tokenizer.

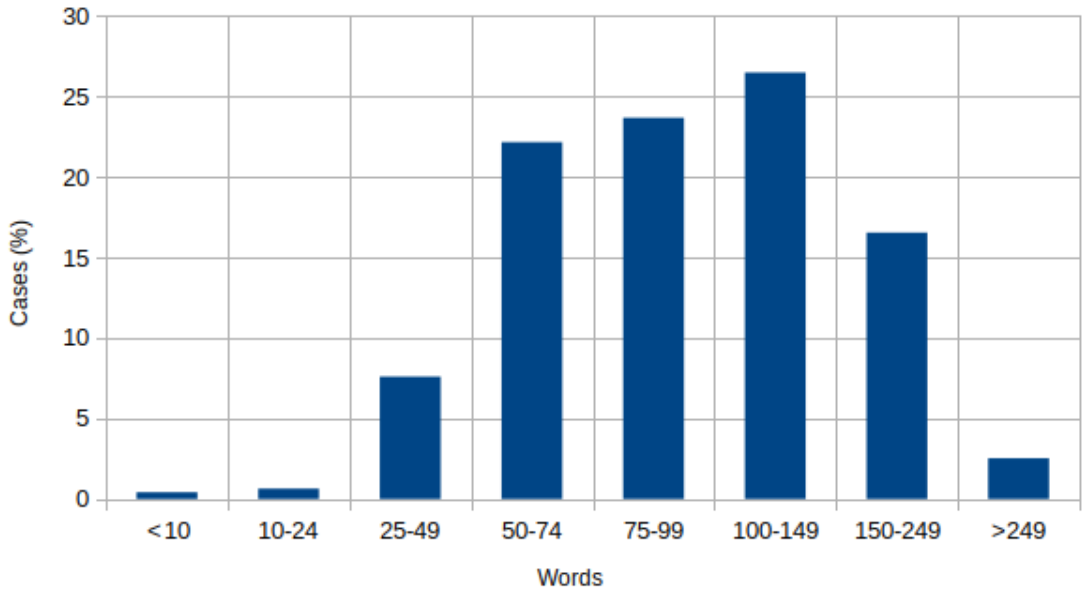

Figure 9: Word distribution for generated summaries

4.2.1 Qualitative evaluation

In our manual evaluation we could see that tokenization improved significantly. An example of how this improved sentence splitting can be found in Appendix B, where

you can see that abbreviations such as 'WW.' and 'art.' are now handled properly. Because of this, the readability improved of the summary. We also noticed a small increase in quality of the summaries when looking at relevance. One reason for this is the increase in training data, which has improved the way the model generates and finds important sentences.

Appendix E discusses two examples of generated summaries from our test set in more detail. In general, we observed that the model does not introduce many novel sentences. Still, it showed good results for rewriting sentences and removing redundant details from the case. When important facts (e.g numbers and dates) showed up, the model immediately switched to a more extractive approach to make sure no errors were made. However, sometimes the model recognized words as not important, probably because it has not seen them before, as discussed in the first example in Appendix E. This lead to sentences being cut off too fast. Also, the summary did not always include all elements that are needed in a summary (as discussed in Section 3.1: background, considerations and judgement), as most of the time only two of these elements were included.

4.3 Specific evaluation

We have used the pre-trained model from the previous phase to evaluate on certain parts of our test set. First, we evaluated on specific categories of cases, as can be seen in Table 11. Administrative Law performs best on ROUGE scores, while Civil Rights is among the worst performers. As can be seen in Table 5, over half of all cases on Rechtspraak are related to Administrative Law. Thus, the model trains on more cases of this type and better learns how to summarize them. Still, Civil Rights is also often present on Rechtspraak, close to 30% of all cases. For all types, the model generated summaries of around 6 sentences, with an average of 120 words.

Type	**Cases**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Administrative Law	3116	39.26 (38.82, 39.74)	18.05 (17.49, 18.64)	35.99 (35.54, 36.49)
Civil Rights	1905	32.83 (32.30, 33.36)	10.72 (10.24, 11.26)	29.46 (28.96, 29.98)
Criminal Law	2198	37.54 (36.94, 38.12)	17.48 (16.76, 18.18)	34.73 (34.13, 35.34)
Tax Law	786	36.46 (35.63, 37.32)	13.46 (12.59, 14.42)	32.83 (32.03, 33.70)

Table 11: F_1 ROUGE scores for category specific evaluation based on a 95% confidence interval

We have also used our pre-trained model from the previous phase to evaluate on different date periods in our test set. Table 12 shows the ROUGE scores for different year periods. In general, the model seems to perform best on cases between 2001 and 2008, while performing worst on cases from the last decade. An average of 6 sentences and 120 were generated by the model for each summary.

Period	Cases	ROUGE-1	ROUGE-2	ROUGE-L
1970-2000	50	38.91 (35.10, 43.05)	16.64 (12.35, 21.37)	35.23 (31.39, 39.32)
2001-2008	1761	38.86 (38.19, 39.53)	18.49 (17.61, 19.38)	35.63 (34.93, 36.32)
2009-2018	5341	36.59 (36.23, 36.94)	15.34 (14.92, 15.74)	33.47 (33.10, 33.83)

Table 12: F_1 ROUGE scores for date evaluation based on a 95% confidence interval

4.4 Abstractiveness and technical evaluation

First, this section discusses the abstractiveness of this model and compares the results with state-of-the-art models that used the CNN/Daily Mail dataset. After this, it will continue with an evaluation of how the ROUGE, abstractiveness and training time changes over the amount of cases. For this evaluation, we have used the same hyperparameters as phase 2 (improved model).

4.4.1 Abstractiveness

In this phase, we discuss the abstractiveness of our model (from phase 2). Following the approach of See et al. (2017), we compute an abstractiveness score as the ratio of novel n-grams in the generated summary that are not present in the original text of the input document (court judgement document). We compare these scores with the abstractiveness of the reference summaries from Rechtspraak, as well as with abstractiveness scores from the model of Chen & Bansal (2018) and See et al. (2017) on the CNN/Daily Mail dataset.

Model	Dataset	1-gram	2-gram	3-gram	4-gram
See et al. (2017)	CNN/Daily Mail	0.1	2.2	6.0	9.7
Chen & Bansal (2018)	CNN/Daily Mail	0.3	10.0	21.7	31.6
Reference summaries	CNN/Daily Mail	10.8	47.5	68.2	78.2
Chen & Bansal (2018)	**Rechtspraak**	**1.9**	**8.4**	**13.2**	**16.5**
Reference summaries	Rechtspraak	5.4	26.7	39.5	45.7

Table 13: Abstractiveness: the ratio (%) of novel n-gram counts of summaries compared to the input text. The model and dataset used in this research is highlighted in bold

The results are shown in Table 13. Using the model of Chen & Bansal (2018), we achieve a higher abstractiveness compared to the model of See et al. (2017) on the CNN/Daily Mail dataset. Still, our generated summaries generate less unique n-grams compared to the generated summaries by Chen & Bansal (2018) on the CNN/Daily Mail dataset. We argue that this occurs because the reference summaries of CNN/Daily Mail generate more unique n-grams than the reference summaries of Rechtspraak. As our model is trained on the reference summaries, which use relatively more n-grams of the input document, one can understand that our model shows the same behaviour.

In our manual evaluation, we could see that the model extracts many sentences from the input document itself. When looking at sentences with similar 4-grams, we observed that the model actually used much larger n-grams from the text. However, the model did rewrite and shortened many sentences, thus improving the readability of the text. In addition, redundant information from sentences was removed properly, which made sentences get to the point more quicker. However, we did note that the model occasionally tends to remove necessary facts and details from sentences, which are needed to understand the case.

Figure 10 shows the abstractiveness of our model compared to the input document, when changing the number of cases. One can see that the generated summaries get less abstractive over time, with a peak around 15K & 20K cases. One reason for this is that the model, with more training data, learns to more often rewrite sentences from the input document, instead of generating new sentence using the vocabulary.

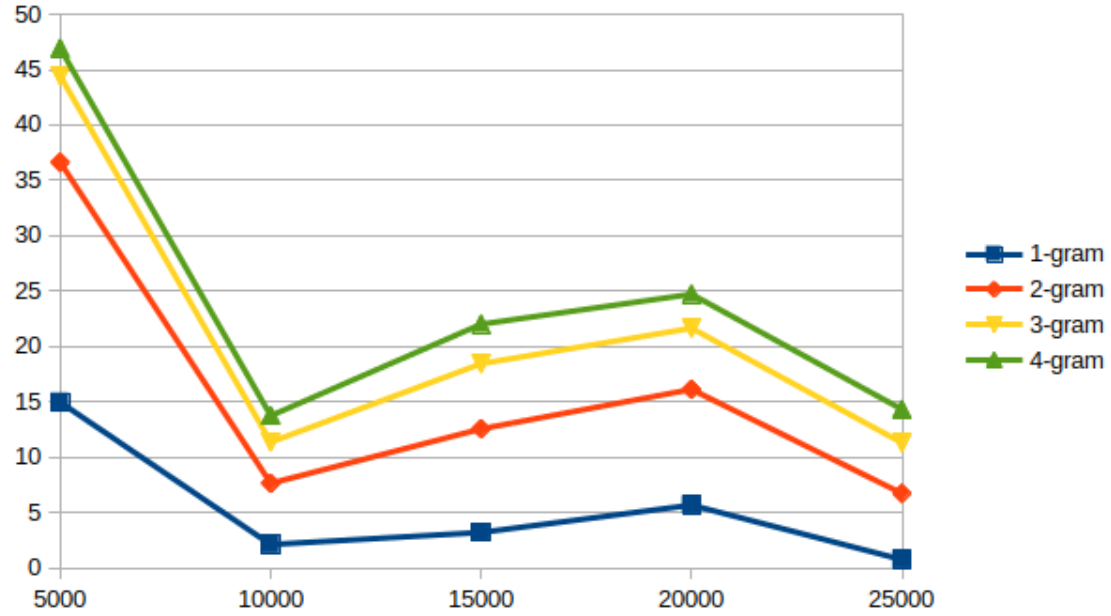

Figure 10: Ratio of novel n-gram counts of summaries compared to the input text and number of cases

Table 14 shows the abstractiveness of our model compared to the reference summaries of Rechtspraak. Here, we compute the abstractiveness as the ratio of novel n-grams in the generated summary that are not present in the reference summary. One can see that our model generates very different summaries from the reference summaries. For example, only 10% of 4-grams in our generated summaries are identical to 4-grams in the reference summary.

1-gram	**2-gram**	**3-gram**	**4-gram**
64.4	84.39	88.71	90.18

Table 14: Abstractiveness: the ratio (%) of novel n-gram counts in our generated summaries compared to the reference summary

In Figure 11 the abstractivess of our model compared to the reference summary over the number of cases is illustrated. Here, you can see that the model uses more sentences from the reference summary when more training data is available. This behaviour is expected, as the model trains on these reference summaries, and thus tends to better 'replicate' these summaries with more data.

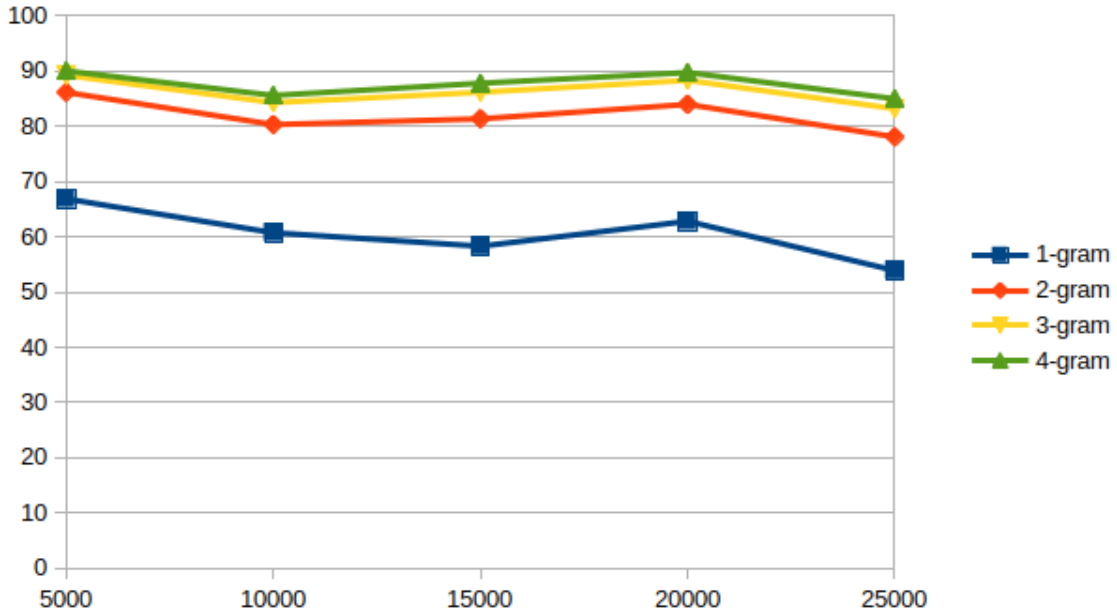

Figure 11: Ratio of novel n-gram counts of summaries compared to the reference summary and number of cases

4.4.2 Technical evaluation

This section describes how the training time and ROUGE score changes over the number of cases. Also, we discuss the relationship between ROUGE and abstractiveness. Figure 12 displays how the training time increases over time. As expected, the training time increases with the number of cases, but there is no linear relationship.

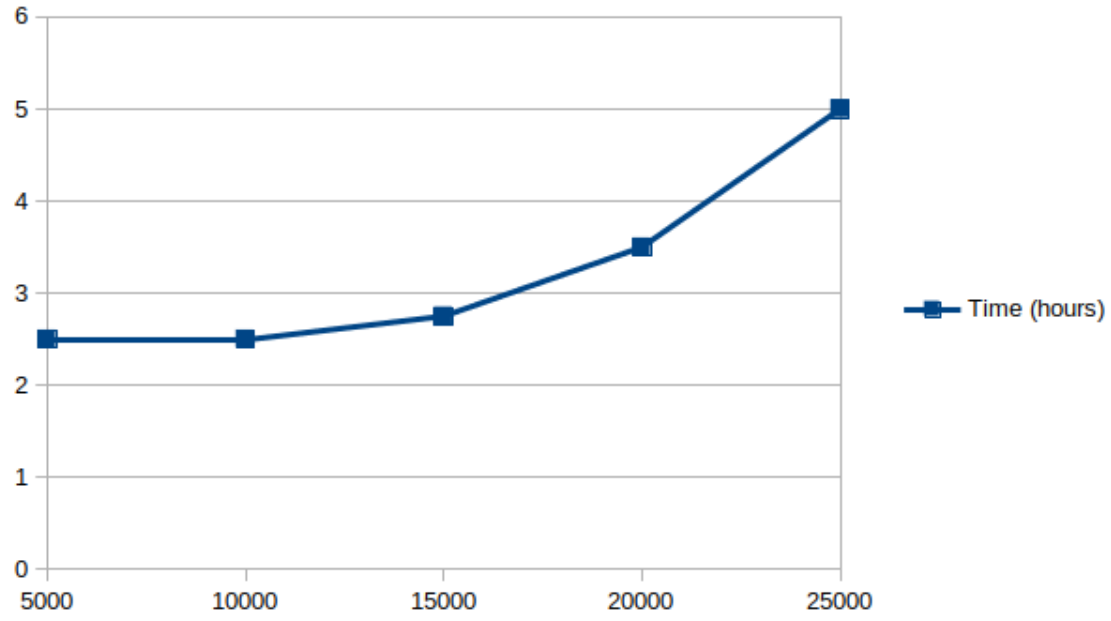

Figure 12: Training time and number of cases

Figure 13 illustrates the relationship between ROUGE-L scores and the number of cases. We have decided to look at the ROUGE-L score, because ROUGE-1 and ROUGE-2 scores suffer from issues where sentences have the same score but different meanings Schluter (2017). What we can see is that the F_1 score stays quite constant after 5K cases, but precision tends to increase over time, with recall decreasing; the model tends to be more careful in selecting sentences.

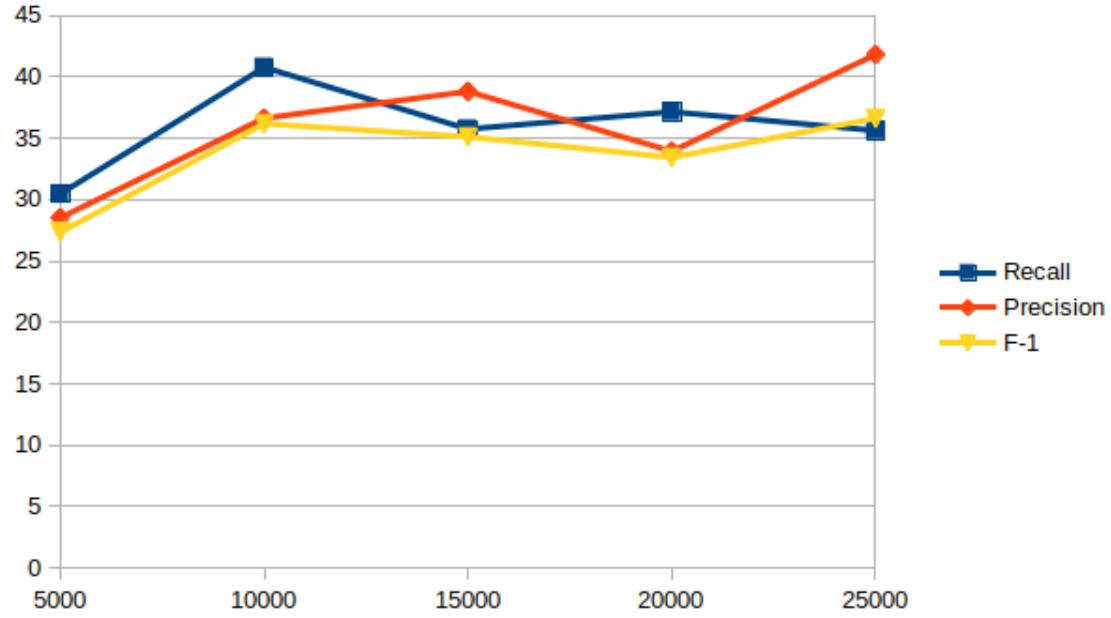

Figure 13: ROUGE-L score and number of cases

Figure 14 shows both the ROUGE-L F_1 score and the ratio of novel n-grams compared to the reference summary in one single graph. What we can observe, is that

the abstractiveness and ROUGE score have an inverse relationship. The more abstractive a summary becomes, the more the ROUGE score decreases. One can argue whether ROUGE should be the standard choice for summary evaluation, especially as models get more and more abstractive.

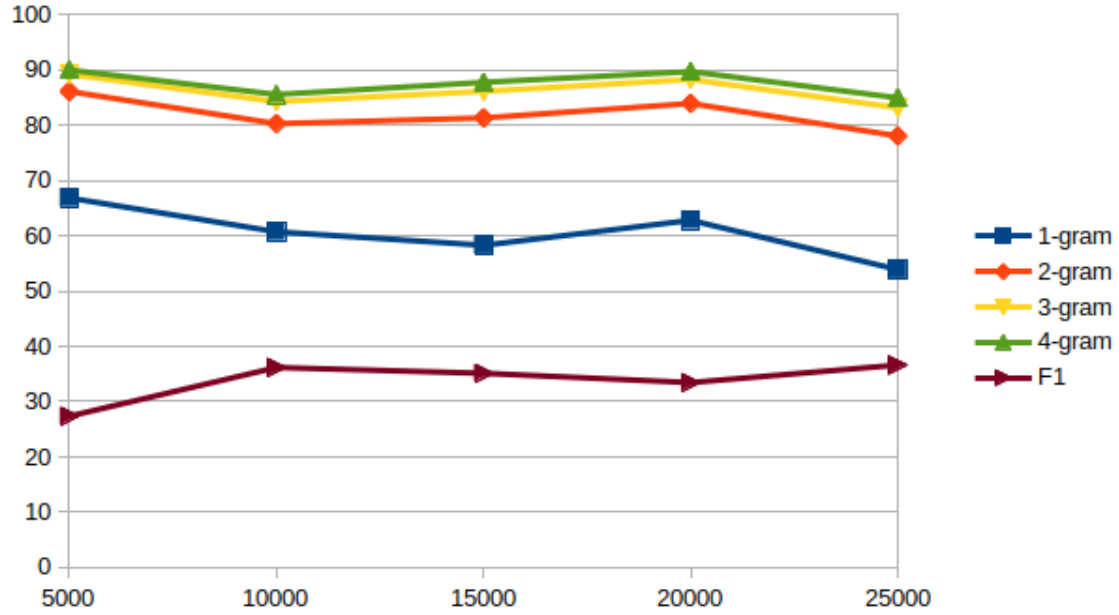

Figure 14: ROUGE-L F_1 score and ratio of novel n-grams compared to the reference summary over number of cases

4.5 Human evaluation

As discussed before, a total of five generated summaries and reference summaries were evaluated on both relevance and readability by two subjects in the first evaluation phase. Table 15 described the results of the first evaluation phase. Our data was tested for normality using the Shapiro–Wilk (Shapiro & Wilk, 1965) test and follows a normal distribution. Relevance and readability of the generated summaries (n=10) averaged 4.60 ± 2.12 and 5.55 ± 1.67 respectively. On the other hand, relevance and readability of reference summaries (n=10) averaged 6.65 ± 1.63 and 7.00 ± 1.63 respectively.

Summary	**Relevance**	**Readability**
Generated summary	4.60 ± 2.12	5.55 ± 1.67
Reference summary	6.65 ± 1.63	7.00 ± 1.63

Table 15: Results of the human evaluation experiment for measuring relevance (0-10) and readability (0-10) of summaries, reported using the mean and its standard deviation ($n = 10$)

We define the null hypothesis, H_0, as there being no difference in the mean (either relevance or readability) between the generated summary and the reference summary. The alternative hypothesis, H_a, defines a difference in the mean (either

relevance or redability) between the generated summary and the reference summary. Results display a significant increase in relevance (2.05 ± 2.83) for reference summaries (paired sample t-test, $p = 0.048$, 95% confidence interval), rejecting the null hypothesis. Readability shows an increase (1.45 ± 2.34) for reference summaries, which appears to be insignificant (paired sample t-test, $p = 0.082$, 95% confidence interval), thus accepting the null hypothesis.

Table 16 describes the results of the second phase of our human evaluation. The objective was to evaluate which summary the subject would prefer, if he were a legal expert at a consulting company that give advice on legal related issues. Here, a reference summary containing only keywords was used for comparison. All subjects prefered to use the generated summary instead of the reference summary that only contained keyword. Students explicitly noted that they disliked a legal case to be summarised using only keywords, as this approach is much too abstract.

Summary	**Preference**
Generated summary	10
Reference summary (keywords)	0

Table 16: Results of our human evaluation experiment for comparing our generated summary with a reference summary containing only keywords (n=10)

The main issue found in the generated summaries regarding relevance was that not all elements needed in the summary (background, considerations and judgement) were always present together. For example, some summaries incuded the facts and the judgements of a case, but failed to explain the considerations. Here, reference summaries did a better job, but also tended to miss fundamental details of cases. Still, the students argued that the text that was in the generated summaries did contain important information about the case.

An observation on generated summaries by the law students regarding readability was that sentences in the generated summary occasionally had sentences that were not grammatically correct or ended in a weird way. Also, the order of sentences was criticized in both generated and reference summaries, as some summaries started with the judgement and ended with background information about the case. Appendix F.1 discusses one of the cases in more detail.

5 Discussion

In this thesis we discussed the application of abstractive summarization in the legal field. This idea gained a lot of criticism in literature, as researched argued that an abstract may be less accurate and can lead to misinterpretations of a judge's intent. Still, we proposed that due to more data being available, improved hardware and matured algorithms, abstractive models now have enough potential to be useful in the legal field. This study challenged two gaps in current research on text summarization. First, no research to date has made use of an abstractive summarization model in the legal context. In addition, no work on abstractive summarization has

evaluated their model on documents of the size and structure of a legal document, which are considerable larger than documents used in current datasets.

A dataset containing over 400K Dutch verdicts was used to train the abstractive summarization model introduced by Chen & Bansal (2018). A five-phased approach was followed to evaluate generated summaries based on ROUGE, abstractiveness and through a human evaluation experiment using law graduates. Our experiments report an F_1 score of 37.24 (ROUGE-1), 16.20 (ROUGE-2) and 34.07 (ROUGE-L) using a dataset of 47K cases, comparable to state-of-the-art results achieved on the CNN/Daily Mail dataset. In our qualitative evaluation, we discussed that the model did not introduce many novel n-grams, but showed good performance for rewriting sentences. Still, the model cut sentences off too early and failed to always include all elements (background, considerations and judgement) in the summary. The summary did include important facts of the case, but often went too deep into detailed specifics.

Our model achieved a higher abstractiveness compared to the input document than the model of See et al. (2017) on the CNN/Daily Mail dataset, but failed to beat the model of Chen & Bansal (2018). We argue that this occured because reference summaries of CNN/DailyMail are generally more abstractiveness than reference summaries of Rechtspraak that we trained on. Also, we noticed that the strength of our model is that it rewrites the long and redundant sentences found in legal text to much shorter ones. This is supported by the observation that our model tends to use less new words with more data, shortening and rewriting sentences instead. In our qualitative evaluation, we noted that our model generates very different summaries (novel n-grams) than the reference summaries from Rechtspraak. Due to the inverse relationship of the abstractiveness and ROUGE score of a document, we doubt if ROUGE is still a good metric that should be used in evaluating abstractive summarization.

A human evaluation study using law students was conducted to evaluate both generated and reference summaries on both relevance (0-10) and readability (0-10). Results show a significant difference in relevance between reference summaries (6.7) and generated summaries (4.6). On the other hand, the difference in readability between reference summaries (7.0) and generated summaries (5.6) appear to be insignificant. The students observed that the main issue in generated summaries was that all elements needed in a summary (background, considerations and judgement) were always present together. For readability, sentences often ended in a weird way or were not grammatically correct. Still, the students noted that the generated summaries did contain key information about the case and preferred it to using a reference summary with only keywords.

Through this research, we argue that there is a lot of potential for abstractive summarization in the legal field. The model of Chen proved to be good in rewriting the long and redundant sentences found in legal text to shorter ones. Also, we have showed that the abstractive model can effectively be applied on documents of large sizes, as current research is limited to news articles. By rewriting the model of Chen & Bansal (2018) to use the state-of-the-art artificial intelligence frameworks, we have also contributed to the general field of text summarization.

Still, the summaries tend to not always include the three core elements that are

needed in a summary. Also, our qualitative evaluation and human evaluation showed that our model cut sentences off too early, leading to grammatically incorrect sentences. We argue that the generated summaries are not of the quality as the reference summaries yet, but are a good replacement for summaries that only contain a set of keywords.

For improving readability, a parser can be implemented in the decoding function that can give a signal when a sentence is cut off too early, giving this sentence a lower score in beam search. Also, post-processing can fix some problems regarding nouns, as the model did not always use these correctly when rewriting sentences.

For improving relevance, an implementation of a neural network that can identify the three core elements needed in a summary can prove useful. One can implement a clustering algorithm such as K-means to find diverse topics in the text, and then identify the most important sences in these clusters. Also, a better embedding library such as BERT Devlin et al. (2018) can help identify representations of words in different contexts, generating multiple embeddings for a single word. Finally, a larger human evaluation using law experts can be conducted, as we were limited to only using law students.

References

Al-Sabahi, K., Zuping, Z., & Kang, Y. (2018). Bidirectional Attentional Encoder-Decoder Model and Bidirectional Beam Search for Abstractive Summarization. *arXiv:1809.06662 [cs]*. arXiv: 1809.06662.

Alschner, W. & Skougarevskiy, D. (2017). Towards an Automated Production of Legal Texts Using Recurrent Neural Networks. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, (pp. 229–232)., New York, NY, USA. ACM. event-place: London, United Kingdom.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. arXiv: 1409.0473.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

Cao, Z., Li, W., Li, S., Wei, F., & Li, Y. (2016). AttSum: Joint Learning of Focusing and Summarization with Neural Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (pp. 547–556)., Osaka, Japan. The COLING 2016 Organizing Committee.

Cao, Z., Wei, F., Dong, L., Li, S., & Zhou, M. (2015). Ranking with Recursive Neural Networks and Its Application to Multi-document Summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, (pp. 2153–2159)., Austin, Texas. AAAI Press.

Carbonell, J. & Goldstein, J. (1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, (pp. 335–336)., New York, NY, USA. ACM.

Chalkidis, I., Androutsopoulos, I., & Michos, A. (2017). Extracting Contract Elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, (pp. 19–28)., New York, NY, USA. ACM. event-place: London, United Kingdom.

Chen, Y.-C. & Bansal, M. (2018). Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. *arXiv:1805.11080 [cs]*. arXiv: 1805.11080.

Cheng, J. & Lapata, M. (2016). Neural Summarization by Extracting Sentences and Words. *arXiv:1603.07252 [cs]*. arXiv: 1603.07252.

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 93–98)., San Diego, California. Association for Computational Linguistics.

Compton, P. & Jansen, R. (1990). A philosophical basis for knowledge acquisition. *Knowledge Acquisition*, *2*(3), 241–258.

Conrad, J. G. & Al-Kofahi, K. (2017). Scenario Analytics: Analyzing Jury Verdicts to Evaluate Legal Case Outcomes. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, (pp. 29–37)., New York, NY, USA. ACM. event-place: London, United Kingdom.

Cumbley, R. & Church, P. (2013). Is "Big Data" creepy? *Computer Law & Security Review*, *29*(5), 601–609.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805.

Elnaggar, A., Gebendorfer, C., Glaser, I., & Matthes, F. (2018). Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification. *arXiv:1810.07513 [cs, stat]*. arXiv: 1810.07513.

Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Fan, A., Grangier, D., & Auli, M. (2017). Controllable Abstractive Summarization. *arXiv:1711.05217 [cs]*. arXiv: 1711.05217.

Farzindar, A. & Lapalme, G. (2004). Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles. In Marie-Francine Moens, S. S. (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, (pp. 27–34)., Barcelona, Spain. Association for Computational Linguistics.

Ferreira, R., de Souza Cabral, L., Lins, R. D., Pereira e Silva, G., Freitas, F., Cavalcanti, G. D. C., Lima, R., Simske, S. J., & Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, *40*(14), 5755–5764.

Galgani, F., Compton, P., & Hoffmann, A. (2012). Combining Different Summarization Techniques for Legal Text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, (pp. 115–123)., Stroudsburg, PA, USA. Association for Computational Linguistics.

Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-Up Abstractive Summarization. *arXiv:1808.10792 [cs]*. arXiv: 1808.10792.

Grover, C., Hachey, B., & Korycinski, C. (2003). Summarising Legal Texts: Sentential Tense and Argumentative Roles. In *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5*, HLT-NAACL-DUC '03, (pp. 33–40)., Stroudsburg, PA, USA. Association for Computational Linguistics.

Hachey, B. & Grover, C. (2006). Extractive summarisation of legal texts. *Artificial Intelligence and Law*, *14*(4), 305–345.

Hahn, U. & Mani, I. (2000). The challenges of automatic summarization. *Computer*, *33*(11), 29–36.

Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., & Uszkoreit, J. (2017). One Model To Learn Them All. *arXiv:1706.05137 [cs, stat]*. arXiv: 1706.05137.

Keneshloo, Y., Shi, T., Ramakrishnan, N., & Reddy, C. K. (2018). Deep Reinforcement Learning For Sequence to Sequence Models. *arXiv:1805.09461 [cs, stat]*. arXiv: 1805.09461.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-Thought Vectors. *arXiv:1506.06726 [cs]*. arXiv: 1506.06726.

Li, P., Lam, W., Bing, L., & Wang, Z. (2017). Deep Recurrent Generative Decoder for Abstractive Text Summarization. *arXiv:1708.00625 [cs]*. arXiv: 1708.00625.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens, S. S. (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, (pp. 74–81)., Barcelona, Spain. Association for Computational Linguistics.

Lu, C., Hsieh, C., Chang, C., & Yang, C. (2013). An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation. In *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, (pp. 463–468).

Mani, I. (2001). *Automatic Summarization*. John Benjamins.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 55–60)., Baltimore, Maryland. Association for Computational Linguistics.

MarketsandMarkets (2018). Cloud Storage Market: Global Forecast until 2022.

Merchant, K. & Pande, Y. (2018). NLP Based Latent Semantic Analysis for Legal Text Summarization. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, (pp. 1803–1807).

Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing Order into Texts. In Lin, D. & Wu, D. (Eds.), *Proceedings of EMNLP 2004*, (pp. 404–411)., Barcelona, Spain. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, (pp. 3111–3119)., USA. Curran Associates Inc.

Moens, M.-F. & de Busser, R. (2002). First steps in building a model for the retrieval of court decisions. *International Journal of Human-Computer Studies*, *57*(5), 429–446.

Nallapati, R., Zhai, F., & Zhou, B. (2016). SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. *arXiv:1611.04230 [cs]*. arXiv: 1611.04230.

Nallapati, R., Zhou, B., & Ma, M. (2016). Classify or Select: Neural Architectures for Extractive Document Summarization. *arXiv:1611.04244 [cs]*. arXiv: 1611.04244.

Nallapati, R., Zhou, B., Santos, C. N. d., Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *arXiv:1602.06023 [cs]*. arXiv: 1602.06023.

Nenkova, A. & Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Nomoto, T. & Matsumoto, Y. (2001). A new approach to unsupervised text summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, (pp. 26–34)., New Orleans, Louisiana, United States. ACM Press.

Padmakumar, A. & Saran, A. (2016). Unsupervised Text Summarization Using Sentence Embeddings.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*, volume 1 of *1*. Stanford.

Paulus, R., Xiong, C., & Socher, R. (2017). A Deep Reinforced Model for Abstractive Summarization. *arXiv:1705.04304 [cs]*. arXiv: 1705.04304.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 1532–1543)., Doha, Qatar. Association for Computational Linguistics.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the Special Issue on Summarization. *Computational Linguistics*, *28*(4), 399–408.

Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. *arXiv:1509.00685 [cs]*. arXiv: 1509.00685.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Sankaran, B., Mi, H., Al-Onaizan, Y., & Ittycheriah, A. (2016). Temporal Attention Model for Neural Machine Translation. *arXiv:1608.02927 [cs]*. arXiv: 1608.02927.

Saravanan, M. & Ravindran, B. (2010). Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, *18*(1), 45–76.

Saravanan, M., Ravindran, B., & Raman, S. (2006). Improving Legal Document Summarization Using Graphical Models. In *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, (pp. 51–60)., Amsterdam, The Netherlands, The Netherlands. IOS Press.

Schluter, N. (2017). The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, (pp. 41–45)., Valencia, Spain. Association for Computational Linguistics.

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. *arXiv:1704.04368 [cs]*. arXiv: 1704.04368.

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Sulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & van Genabith, J. (2017). Exploring the Use of Text Classification in the Legal Domain. *arXiv:1710.09306 [cs]*. arXiv: 1710.09306.

Sulea, O.-M., Zampieri, M., Vela, M., & van Genabith, J. (2017). Predicting the Law Area and Decisions of French Supreme Court Cases. *arXiv:1708.01681 [cs]*. arXiv: 1708.01681.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 3104–3112). Curran Associates, Inc.

van Gompel, M., van der Sloot, K., & van den Bosch, A. (2012). Ucto: Unicode Tokeniser. (ILK Technical Report).

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 2692–2700). Curran Associates, Inc.

Williams, R. J. & Zipser, D. (1989). A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, *1*(2), 270–280.

Yousfi-Monod, M., Farzindar, A., & Lapalme, G. (2010). Supervised Machine Learning for Summarizing Legal Documents. In Farzindar, A. & Kešelj, V. (Eds.), *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, (pp. 51–62). Springer Berlin Heidelberg.

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., & Zhao, T. (2018). Neural Document Summarization by Jointly Learning to Score and Select Sentences. *arXiv:1807.02305 [cs]*. arXiv: 1807.02305.

A Chosen models

This section discusses the models by Paulus et al. (2017), Chen & Bansal (2018) and Fan et al. (2017) more in-depth and describes their approach in more detail.

A.1 Reinforced Learning

Paulus et al. (2017) presented an abstractive summarization method that was designed to address the repeating phrase problem. This problem was illustrated in an analysis by Nallapati et al. (2016) where attentional encoder-decoder models often generated unnatural summaries consisting of repeated phrases. Their model is based on the encoder-decoder network by Sutskever et al. (2014) and makes use of a word embedding matrix for bot input and output embeddings.

At each decoding step, an intra-temporal attention function (Sankaran, Mi, Al-Onaizan & Ittycheriah, 2016) is used that evaluates over the input sequence and looks at both the decoder's own hidden state as well as the previously generated word. This ensures that the model uses different parts of the input sequence and does not uses the same part of the input on different decoding steps. Still, the decoder can generate repeated phrases based on its own hidden states. To prevent this, the model incorporates information about previous decoding steps which allows the model to make more accurate predictions, even if the repeated phrase was decoded many steps way.

The decoder uses a switch function that decides at each step to use either a token generator or a pointer mechanism to copy rare or unknown words from the input Nallapati et al. (2016). This token generator is based on a soft-max layer over a target vocabulary to generate words. The words are generated using word embeddings (GloVe) using an input vocabulary size of 150K tokens, and the output vocabulary to 50K tokens by selecting the most frequent tokens.

Traditionally, decoder RNNs make use of a teacher forcing algorithm (Williams & Zipser, 1989) that minimizes a maximum-likelihood loss at each decoding step. However, using this algorithm does not always produce the best results on evaluation metrics such as ROUGE. The first reason comes from the fact that during training time, the network uses its own generated sequences as input for the next step. This becomes a problem at test time, as (wrongly) generated words are feeded back to the network, which easily accumulates errors over time in long documents. Another reason is that there are so many different ways to write a summary. ROUGE metrics have some flexibility due to sentence ordering and producing paraphrases, bt the maximum-likelihood objective does not.

One solution to solve the above discussed problem is to find an objective function that maximizes a discrete measurement such as ROUGE. Still, an issue could be that optimizing for this metric does not guarantee improved quality and readability of the summary. Paulus et al. (2017) have decided to implement a mixed training objective function that combines both the teacher forcing algorithm (ML) and a function that maximizes a reinforcement training objective based on ROUGE (RL).

In their evaluation, the authors used various models that variated between seperately using ML and RL, to combing both of these functions. Expectedly, the RL algorithm was the top performer with a ROUGE-1 of 41.16% (+2%) and ROUGE-L of 39.08%(+3%) compared to the other variations. To ensure that this increase in ROUGE was also followed by an increase in human readability and quality, a human evaluation was conducted as well. One hundred random test examples were selected from the dataset and for each example the authors showed the article, the ground truth summary and the summaries generated by all models. 5 different evaluators from Amazon Mechanical Turk, who did not know which summaries came from which model or which one was the ground truth, had to assign a score from 1 to 10 based on relevance and readability.

Results showed that, despite the RL algorithm achieving the highest ROUGE-1 and ROUGE-L score, it produced the least readable (4.18) and relevant (6.32) summaries. The most common issue was that there were a lot of short and truncated sentences towards the end of the summary. The top performer was the ML+RL model, achieving a readability and relevance of respectivally 7.04 and 7.45, while the ML model had a 6.76 and 7.14.

A.2 Fast Abstractive Rewriting

Chen & Bansal (2018) proposed a model that first selects important sentences and then rewrites them abstractively. First, sentences are represented using a temporal convolutional model and words are converted to a distributed vector representation by using word embeddings. Sequences of word vectos are fed through the layers of the model to capture the dependencies of nearby words.

Selecting sentences from the above sentence representations is done by training a pointer network based on a set of features (Vinyals, Fortunato & Jaitly, 2015). From these extracted sentences, an abstractive model compresses and paraphrases these sentences in order to create a concise summary sentence. The encoder-decoder structure by Bahdanau et al. (2014) is used and a copy mechanism See et al. (2017) is added to help out with out-of-vocabulary words. The objective function is standard and minimizes a maximum-likelihood loss at each decoding step. Similar to Paulus et al. (2017), reinforced learning techniques are used based on ROUGE. If the abstractive model rewrites a sentences and the ROUGE match is high, the action is encouraged. On the other hand, if the ROUGE score is low, the model discourages this action. In this reinforced learning phase, an additional trainable parameter 'End-Of-Extraction' is added, that rewards the agent for finding the correct number of sentences for the summary. The words are generated using word embeddings (word2vec) using an output vocabulary of 30K tokens by selecting the most frequent ones.

In their evaluation, the authors used several variations of the proposal model. Results showed that the reinforced learning algorithm has significantly improved the model on all ROUGE-scores. Similar to Paulus et al. (2017), the authors also performed a human evaluation using Amazon Mechanical Turk. Using 100 samples from the CNN/Daily Mail dataset, three human evaluaters were asked to rank summaries on relevance and readability, comparing their model and the pointer model

of (See et al., 2017). The model of Chen & Bansal (2018) outperformed the pointer model by 15% on relevance and 5% on readability. Also, results showed that their model creates many novel bigrams and trigrams that are not present in the input document, indicating a high abstractiveness and the ability to create novel sentences.

A.3 Controllable Abstractive Summarization

Fan et al. (2017) designed an abstractive summarization model with a mechanism to regard user preferences such as the desired length, style, entities the mdoel should focus on and how much of the document the user has already read. As opposed to Chen & Bansal (2018) and Paulus et al. (2017), the authors do not employ a pointer network to copy rare entities from the input. Instead, they rely on sub-word tokenization and weight sharing, which adds less complexity to the model. The model enables the user to control length by quantizing summary length into descrete bins with a size range. The input vocabulary and training documents are then configured in such a way so that training is optimized based on this discrete length variable.

This model enables the user to select a certain entity (people or location) that the summary should be focused on. At training time, each document is filled with markers that refer to an entity from the ground-truth summary. Another interesting aspect is that users can specify a preferred source style for a summary, such as a newspaper or magazine. A special maker token is introduced to express the desired source data.

Finally, the model allows readers who only read the first few paragraphs to only have the rest of the text summarized. To enable remainder document summarization, the authors employ full text summarization and remove sentences aligned before the point that the user specifies.

For their human evaluation, a study using Amazon Mechanical Turk is conducted using 500 articles from the test set that were evaluated by five raters compared to the study of See et al. (2017). Results showd that 59% of raters prefered the summaries of Fan et al. (2017).

B Examples of sentence splitting improvements

Before:

- Dit houdt in dat van betrokkene op die dag moet worden gezegd dat hij werkloos was, zodat ook op die dag het bepaalde in art.
- 24, lid 1 sub b, WW op hem van toepassing kon zijn.

After

- Dit houdt in dat van betrokkene op die dag moet worden gezegd dat hij werkloos was, zodat ook op die dag het bepaalde in art. 24, lid 1 sub b, WW op hem van toepassing kon zijn.

Before

- Het college heeft zich laten vertegenwoordigen door J.W.B.
- van den Berg, werkzaam bij de gemeente Zeist.

After

- Het college heeft zich laten vertegenwoordigen door J.W.B. van den Berg, werkzaam bij de gemeente Zeist.

Before

- Dit oordeel is door de Raad nadien bij vele uitspraken – waarbij mr.
- De Jonge als gemachtigde optrad – herhaald.

After

- Dit oordeel is door de Raad nadien bij vele uitspraken – waarbij mr. De Jonge als gemachtigde optrad – herhaald.

C AWS Architecture

For the model to compute on a large dataset in a reasonable time, it requires the massively parallel processing power of a GPU. Chen & Bansal (2018) used an AWS instance with an Nvidia Tesla K80 GPU in their research, and still it took over 24 hours for training. Because of this, we have also decided to use a GPU optimized AWS instance for training. This instance (p3.2x large) uses an NVIDIA Tesla V100, which is considerably faster than the GPU used by Chen & Bansal (2018) in their research. In addition, we have made use of a CPU and RAM optimized instance (t2.2x large) for pre-processing, decoding and evaluation. Table 17 shows more information about these two instances. The reason that we have not chosen one instance to work with, is that GPU optimized AWS instances are very expensive, while they are only needed for training the model.

Instance	**CPU**	**GPU**	**RAM**
t2.2x large	Intel Xeon 3.0GHz (8-core)	-	32GB
p3.2x large	Intel Xeon 2.7GHz (8-core)	NVIDIA Tesla V100 SXM2 16GB	61GB

Table 17: AWS instances used for this research

First, we used a t2.2x large AWS instance for setting up our working environment and loading the Rechtspraak data, as shown in Figure 15. Then, we saved this data into an Amazon Machine Image (AMI), so new instances can be started with the same environment.

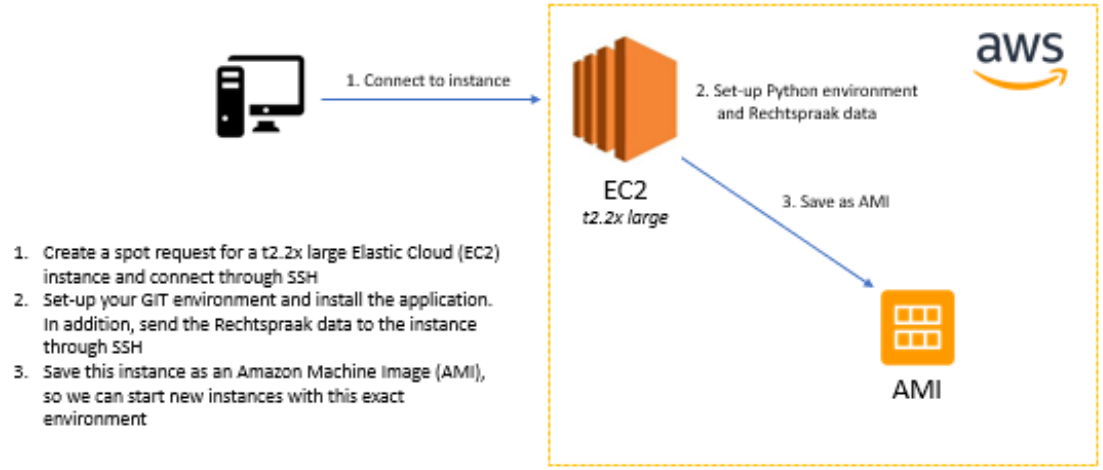

Figure 15: First-time setup

The general workflow in AWS for our pipeline starts with creating a t2.2x large AWS instance for pre-processing our dataset using the AMI created in the first-time setup. Here, we pre-process data based on a set of specified filters as discussed in Section 3.4. We then save this data into an Elastic Block Store (EBS), so we can use the stored dataset in other instances as well.

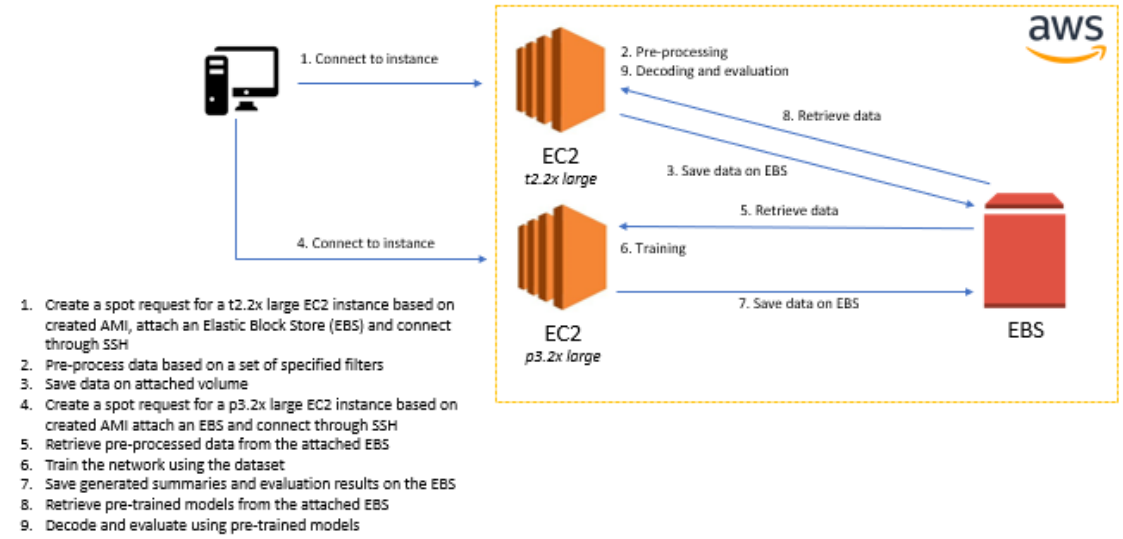

Figure 16: General AWS workflow

For training, we have made use of a GPU optimized p3.2x large instance. After training, the pre-trained models are saved on EBS. Using the t2.2x large AWS instance, the pre-trained models can be used to decode and evaluate summaries.

D Criteria

E Qualitative evaluation

In the example shown in Table 18, the model first gives a very short background description of the case and describes one of the considerations. The description of the background of the case is far too small and the consideration is discussed in far too much detail. Also, the judgement of the case is not discussed. In the first sentence, the main subject of the case (quarantainevoorzieningen levende tweekleppige weekdieren 2007) is removed from the case, likely because the model has not seen this word before and thus does not deem it important. Apart from this mistake, the model does a very good job at rewriting the sentence to a more clearer one. The same thing happens in the second sentence. Also, the 'lidnummer' of the article is skipped. For future work, post processing could help fix these kind of issues. In the third sentence, much unnecessary details are removed. Still, the summary does not include the (important) fact that the 'verweerder' was accused of this case and not that he actually did it. The fourth sentence is taken directly from the article. In general, the summary goes in too much detail on some parts of the case and fails to give a generalized summary. However, this example does show the power of the model of how it can rewrite sentences into much clearer and shorter ones.

Case (ECLI:NL:CBB:2013:212)
. . . . Procesverloop Bij brief van 18 juni 2012 heeft appellante zich gericht tot verweerder met het verzoek om handhavend op te treden tegen [A] B.V. en [B] B.V. wegens (vermeende) overtreding van de bepalingen van de Verordening quarantainevoorzieningen levende tweekleppige weekdieren 2007. Bij brief van 19 juni 2012 heeft verweerder een reactie gegeven. Bij brief van 27 juni 2012 heeft appellante het verzoek om handhavend op te treden herhaald. Bij brief van 9 juli 2012 heeft verweerder opnieuw een reactie gegeven. Op de daarop volgende brieven van appellante van 24 en 27 juli 2012 heeft verweerder bij brief van 3 augustus 2012 gereageerd. . . . Overwegingen . . . 3.1 Niet in geschil is dat appellante heeft verzocht aan verweerder om een tuchtzaak aanhangig te maken wegens (vermeende) overtreding van het verbod van artikel 2, eerste lid, van de Verordening quarantainevoorzieningen levende tweekleppige weekdieren 2007. Appellante heeft verweerder vervolgens in gebreke gesteld omdat verweerder, naar zij stelt, heeft nagelaten te besluiten over het aanhangig maken van een tuchtzaak, en zij heeft daarna op grond van artikel 6:12 van de Awb beroep ingesteld. Daarmee ligt allereerst de vraag voor of beroep op grond van artikel 6:12 van de Awb in dit geval voor appellante openstond. Deze vraag beantwoordt het College ontkennend. In artikel 15 van de Wet tuchtrechtspraak bedrijfsorganisatie 2004 is bepaald dat de voorzitter van verweerder, na constatering van de (vermeende) overtreding, de zaak binnen een redelijke termijn bij het tuchtgerecht aanhangig maakt door middel van een schriftelijke verklaring. . . . .
Reference summary
Betreft beroep tegen niet tijdig nemen van een besluit. Een dergelijke procedure staat niet open tegen het niet aanhangig maken van een tuchtprocedure door de voorzitter van het Productschap Vis. Het doen van een dergelijk verzoek is geen aanvraag in de zin van artikel 1:3 van de Awb. De wet tuchtrechtspraak bedrijfsorganisatie kent eigen bepalingen.
Generated summary
Verzoek om handhavend op te treden. Niet in geschil is dat appellante heeft verzocht aan verweerder om een tuchtzaak aanhangig te maken wegens (vermeende) overtreding van het verbod van artikel 2, eerste. Verweerder heeft nagelaten te besluiten over het aanhangig maken van een tuchtzaak. Daarmee ligt de vraag voor of beroep op artikel 6:12 van de Awb in dit geval voor appellante openstond.

Table 18: Example from the dataset showing the generated summary of our model. The colored (marked) sentences correspond to the sentences that were extracted by our model.

In Table 19, we can see that the summary starts with some background information and considerations of the case. Still, it does not give a clear description of the final judgement. One thing to note, is that the model rewrites some sentences, while others are extracted directly from the case. This is because the case contains a lot of important facts such as date and numbers. As the model wants to make certain to retain these facts in the summary, it is more likely to literally take these sentences from the text, not taking any risk of wrongly rewriting sentences. However, just as in the first case we discussed in this qualitative evaluation, some important facts about the background information of the case ("Regeling GLB-inkomensstein 2006") are not included in the summary.

Case (ECLI:NL:CBB:2013:11)
. . . . Procesverloop Bij besluit van 26 mei 2011 (het primaire besluit) heeft verweerder de bedrijfstoeslag van appellante voor het jaar 2010 vastgesteld op grond van de Regeling GLB-inkomenssteun 2006. Bij besluit van 21 december 2011 heeft verweerder het bezwaar van appellante gedeeltelijk gegrond verklaard. Appellante heeft tegen dat besluit beroep ingesteld. Verweerder heeft een verweerschrift ingediend. Op 27 februari 2013 heeft verweerder het bestreden besluit herzien. Het onderzoek ter zitting heeft plaatsgevonden op 22 maart 2013, waarbij voor partijen hun gemachtigden zijn verschenen. . . . Overwegingen 1. Met haar Gecombineerde Opgave 2010 heeft appellante uitbetaling van haar toeslagrechten aangevraagd. Zij heeft 26 gewaspercelen met een totale oppervlakte van 51.37 ha opgegeven. Bij het primaire besluit heeft verweerder de bedrijfstoeslag vastgesteld op €18.174,06.Verweerder heeft daarbij 2.21 ha van de door appellante opgegeven percelen afgekeurd en een sanctie (extra korting) toegepast. In het besluit van 27 februari 2013 is de afgekeurde oppervlakte lager vastgesteld, de bedrijfstoeslag (verder) verhoogd naar €20.323,31 en de extra korting vervallen. . . . 2.2 Verweerder heeft de subsidiabele oppervlakte van appellantes percelen vastgesteld na een administratieve controle als bedoeld in artikel 28 van Verordening (EG) nr. 1122/2009. **Deze vaststelling heeft plaatsgevonden door de door appellante opgegeven oppervlakte te vergelijken met de referentiepercelen en de luchtfoto van 2011.** De GPS-meting van de percelen 1 en 2 heeft verweerder ingetekend op de luchtfoto 2011 en vergeleken met zijn eigen meting. Daarbij is te zien dat in de GPS-meting niet subsidiabele oppervlaktes, namelijk van een weg aan de zuidzijde van perceel 1 en een hoek en een weg op perceel 2, zijn meegenomen.
Reference summary
Bedrijfstoeslag, GPS-meting
Generated summary
Verweerder heeft de bedrijfstoeslag van appellante voor het jaar 2010 vastgesteld. Verweerder heeft het bezwaar van appellante gedeeltelijk gegrond verklaard. Zij heeft 26 gewaspercelen met een totale oppervlakte van 51.37 ha opgegeven. Bij het primaire besluit heeft verweerder de bedrijfstoeslag vastgesteld op €18.174,06. Verweerder heeft daarbij 2.21 ha van de door appellante opgegeven percelen afgekeurd en een sanctie -(extra korting) toegepast. **Deze vaststelling heeft plaatsgevonden door de door appellante opgegeven oppervlakte te vergelijken met de referentiepercelen en de luchtfoto van 2011.** In het besluit van 27 februari 2013 is de afgekeurde oppervlakte lager vastgesteld, de bedrijfstoeslag (verder) verhoogd naar €20.323,31.

Table 19: Example from the dataset showing the generated summary of our model. The colored (marked) sentences correspond to the sentences that were extracted by our model.

F Human evaluation experiment

This section first evaluates one case of our human evaluation in more detail. Then, the documents containing the cases and questions that were used in our experiment are described. As discussed before, due to the limited size of this document, we have only included one example for each of the two evaluation phases.

F.1 Qualitative evaluation

The case discussed in Table 20 is relatively small compared to other cases. One interesting thing to note here, is that all sentences are extracted from two paragraphs that are at the end of the document. The generated summary does not offer much background information of the case, but immediately gets to the considerations for the judgement. What we can see is that the model does not generate novel sentences, but rewrites long and redundant sentences to much shorter ones. This leads to a grammatical error in sentence five, but in other cases, sentences are still readable. In addition, the most important parts of the sentences are still kept in the summary. Finally, the summary discusses the final judgement of the case and gives a very short and concise answer.

One student noted that the reference summary was not legally relevant, as the summary did not answer the key problem of the case. Also, sentences were too long which made them hard to read. On the other hand, the generated summary gave a lot of (legal) background information, but failed to briefly sum up the relevant facts. Because of this, the summary is difficult to read and does not provide a good overview of the case.

Case (ECLI:NL:CRVB:2009:BH4572)
... 5. De Raad overweegt als volgt. 5.1. Het geschil tussen partijen spitst zich toe op de vraag of appellant in het bestreden besluit een beslissing heeft genomen op het verzoek in het bezwaarschrift om vergoeding van de proceskosten.Onder verwijzing naar de door appellant aangehaalde zinsnede in het bestreden besluit, stelt de Raad vast dat in het bestreden besluit wel een beslissing is genomen op dat verzoek.Aangezien de afwijzing van dat verzoek samenhangt met de ongegrondverklaring van het bezwaar, heeft de rechtbank terecht vastgesteld dat aan appellant geen proceskostenvergoeding in bezwaar toekwam. 5.2. Nu de rechtbank heeft overwogen dat de afwijzing van het verzoek om tegemoetkoming in de ziektekosten bij het bestreden besluit rechtens juist is, is de Raad van oordeel dat de rechtbank op onjuiste grond het beroep gegrond heeft verklaard en het bestreden besluit heeft vernietigd. 6. De aangevallen uitspraak kan dus in zoverre niet in stand blijven evenmin als de veroordeling van appellant in de proceskosten van betrokkene in beroep en de vaststelling van de vergoeding van het griffierecht. Uit oogpunt van duidelijkheid geeft de Raad er de voorkeur aan de aangevallen uitspraak in haar geheel te vernietigen en te doen hetgeen de rechtbank zou behoren te doen. ...
Reference summary (Relevance: 5.5, Readability: 6.5)
De minister heeft in het bestreden besluit een afwijzende beslissing genomen op het verzoek in het bezwaarschrift om vergoeding van de proceskosten. Afwijzing van dat verzoek hangt samen met de ongegrondverklaring van het bezwaar, zodat de rechtbank terecht heeft vastgesteld dat aan betrokkene geen proceskostenvergoeding in bezwaar toekwam. Inhoudelijk was bestreden besluit juist. De rechtbank heeft ten onrechte een proceskostenveroordeling in beroep uitgesproken en een vergoeding van griffierecht vastgesteld.
Generated summary (Relevance: 5.0, Readability: 5.75)
Het geschil tussen partijen spitst zich toe. De raad stelt vast dat in het bestreden besluit wel een beslissing is genomen op dat verzoek. Aangezien de afwijzing van dat verzoek samenhangt met de ongegrondverklaring van het bezwaar, terecht vastgesteld dat aan appellant geen proceskostenvergoeding in bezwaar toekwam. De rechtbank heeft op onjuiste grond het beroep gegrond verklaard. De aangevallen uitspraak kan dus niet in stand blijven evenmin als de veroordeling van appellant in de proceskosten van betrokkene in beroep en de vaststelling van de vergoeding van. Uit oogpunt van duidelijkheid geeft de raad er de voorkeur aan de aangevallen uitspraak in haar geheel te vernietigen.

Table 20: Example from our human evaluation showing the generated summary of our model. The colored (marked) sentences correspond to the sentences that were extracted by our model. The relevance and readability scores are an average of the grade given by both students

F.2 Explanation and cases

Uitleg

Tijd: 2.5-3 uur

Evaluatie A

1. Je krijgt in totaal 5 rechtsuitspraken van Rechtspraak te zien. Elke uitspraak lees je eerst 15 minuten door.
2. Hierna krijg je twee samenvattingen te zien voor elke rechtsuitspraak. Je beoordeelt elke samenvatting op *Relevantie (1-10)* en *Leesbaarheid (1-10).* Ook geef je een korte uitleg (1-3 zinnen) als onderbouwing op je antwoord.
    a. *Relevantie*: zit de juiste informatie in de samenvatting? Worden overbodige details overgeslagen? Geeft de samenvatting een juist overzicht van de uitspraak?
    b. *Leesbaarheid*: Hoe is de overgang tussen zinnen? Is de grammatica correct? Hoe is de lengte van de samenvatting?

Evaluatie B

1. Je krijgt in totaal 5 rechtsuitspraken van Rechtspraak te zien. Elke uitspraak lees je eerst 15 minuten door.
2. Hierna krijg je een gegenereerde samenvatting en een inhoudsindicatie (met alleen een aantal kernwoorden) te zien. Je beoordeelt nu welke van de twee je liever zou willen gebruiken als rechtsadviseur voor een consulting bureau. Ook beoordeel je de gegenereerde samenvatting weer op *Relevantie (1-10)* en *Leesbaarheid (1-10).* Geef ook graag een korte uitleg (1-3 zinnen) als onderbouwing op je antwoord op de relevantie & leesbaarheid vraag.

Experiment

Evaluatie A

Uitspraak 1 (ECLI:NL:CBB:2003:AI1121)

College van Beroep voor het bedrijfsleven

Enkelvoudige kamer voor spoedeisende zaken

No. AWB 03/644 22 juli 2003

13700 Wet tarieven gezondheidszorg

Uitspraak op het verzoek om voorlopige voorziening in de zaak van:
Stichting Riagg Rijnmond Zuid, te Barendrecht

verzoekster,

gemachtigde: mr. E.W.M. Meulemans, advocaat te Zwolle,

tegen

het College tarieven gezondheidszorg, verweerder,

gemachtigde: mr. A.C. de Die, advocaat te 's-Gravenhage.

1. De procedure

Bij tariefbeschikking van 17 april 2003, nr. 120-1740-03-2 heeft verweerder de tarieven voor verzoekster met ingang van 1 januari 2003 vastgesteld.

Op 28 mei 2003 heeft verzoekster daartegen een bezwaarschrift ingediend.

Op 6 juni 2003 heeft verzoekster zich tot de voorzieningenrechter van het College gewend met het verzoek een voorlopige voorziening te treffen ertoe strekkende dat de door verweerder in de bestreden beschikking gemaakte reservering wordt opgeheven en verzoeksters budget voorcalculatorisch zal worden verhoogd met € 3.032920, --.

Op 26 juni 2003 heeft verweerder een schriftelijke reactie op het verzoek om voorlopige voorziening bij de voorzieningenrechter ingediend.

Bij brief van 4 juli 2003 heeft verzoekster een nadere reactie gegeven.

Op 14 juli 2003 hebben zowel verzoekster als verweerder nog een aantal aanvullende producties ingediend.

De voorzieningenrechter heeft het verzoek behandeld ter zitting van 18 juli 2003, waar partijen bij monde van hun gemachtigden hun standpunt nader hebben toegelicht. Aan de zijde van verzoekster zijn tevens verschenen M.W. Reitsma en W.M. Dieleman. Namens verweerder waren tevens aanwezig mr. H.H.M. Debets en mr. J. Schuurman.

2. De grondslag van het geschil en de vaststaande feiten

Ingevolge artikel 11 van de Wet tarieven gezondheidszorg stelt verweerder beleidsregels vast omtrent de hoogte, de opbouw en de wijze van berekening van een tarief of van onderdelen van een tarief.

Bij circulaire van 12 december 2002 heeft verweerder de betrokken instellingen geïnformeerd over de nieuwe beleidsregels 2003 en een aantal onderwerpen met betrekking tot de geestelijke gezondheidszorg (GGZ), waaronder de beëindiging van het systeem van budgetmaximering.

In de voor 2003 vastgestelde Beleidsregel III-726 is ten aanzien van productieafspraken voor de geestelijke gezondheidszorg over 2003 het volgende bepaald.

"2.7 Productieafspraken

De tussen de instelling en het zorgkantoor gemaakte afspraak over de reële productie voor het komende jaar van verpleeg- verzorgingsdagen, deeltijd behandelingen en face-to-face-contacten. Onder reële productie wordt verstaan productie die volgens inschatting feitelijk zal worden gerealiseerd. De basis voor de productieafspraak wordt gevormd door de realisatie van het voorafgaande jaar. Sterke afwijkingen dienen te onderbouwd en worden bij gebreke daarvan niet in het budget verwerkt."

In de voor 2003 vastgestelde Beleidsregel III-765 ("loon- en materiële kosten"), is bepaald:

"2.2 Normbedragen loon- en materiële kosten

(…); voor de productieparameters wordt in het budget vooraf uitgegaan van de productieafspraak die partijen (instelling en zorgkantoor) hebben gemaakt. Hierop wordt nagecalculeerd. Zie hiervoor artikel 4.

(…)
"4. NACALCULATIE OP BASIS VAN GEREALISEERDE PRODUCTIE

Op de productieafspraken wordt nagecalculeerd. Dit houdt in dat het budget wordt verminderd als de productiewaarde van de gerealiseerde productie lager ligt dan de productiewaarde die in het budget is opgenomen. (…)

Zowel in geval van meer- als minder productie dient het zorgkantoor een verklaring af te geven met betrekking tot de juistheid van het volume en de zorgzwaarte van de gerealiseerde productie. De verklaring wordt mede gebaseerd op een steekproef van voldoende omvang naar de individueel gescoorde productie van patiënten/cliënten om een representatief beeld te krijgen van de mate waarin feitelijk geleverde zorg overeenstemt met de geregistreerde zorg. Hiertoe heeft ZN als leidraad aan de zorg kantoren een productieprotocol voor de materiële controle van de gerealiseerde productie opgesteld. Dit als bijlage bij deze beleidsregel opgenomen productieprotocol dient te worden toegepast.

Op 28 februari 2003 hebben verzoekster en het Zorgkantoor Zuid-Hollandse eilanden te Breda het tariefverzoek ingediend dat heeft geleid tot de bestreden tariefbeschikking.

Bij circulaire van 15 april 2003 informeert verweerder de instellingen GGZ over de afhandeling van de budgetten 2003 als volgt.

"PRODUCTIEAFSPRAKEN 2003

Voor 2003 geldt geen budgetmaximum meer. Het uitgangspunt is nu dat verzekerde zorg moet kunnen worden gerealiseerd. De budgetformulieren 2003 geven als voorlopig resultaat een bedrag van 187 miljoen dat extra wordt gevraagd boven het budget van 2002. Als capaciteitwijzigingen buiten beschouwing worden gelaten bedraagt de toename 168 miljoen. De spreiding van de procentuele mutatie van het budget 2003 ten opzichte van het budget van 2002 loopt uiteen van -/- 20% tot + 60%. Het macrogemiddelde komt uit op 6,7%.

Gegeven het feit dat het productieprotocol nog lang niet overal wordt toegepast bestaat onzekerheid over de realiteitswaarde van afgesproken grote niet onderbouwde productiestijgingen. Het op voorhand aanvaarden van de afspraken zou ongewenste effecten kunnen hebben op de recent opgeheven beperkingen van de GGZ. Aangezien het dit jaar mogelijk is om (tot uiterlijk 1 december) aanvullende productieafspraken in te dienen is besloten om voor 2003 de budgetstijging ten opzichte van het budget 2002 op dit moment vooralsnog te beperken tot 10% (exclusief capaciteitsmutaties).

Voor instellingen waarop dit van toepassing is kunnen partijen vervolgens een aanvulling indienen, als deze kan worden onderbouwd door gerealiseerde en door het zorgkantoor gecontroleerde (op basis van protocol) productie van een gedeelte van 2003 of van heel 2002. De aanvulling kan ook bestaan uit het naar beneden aanpassen van de in eerste instantie gemaakte afspraak.

Aangezien het CTG geen veranderingen in de afspraken wil maken wordt de toename in de productie boven 10% voorlopig verwerkt in een totaalbedrag. Bij een latere aanvulling van de productie wordt dit bedrag aangepast. In beginsel moeten deze bedragen voor het einde van het jaar worden weggewerkt door of hogere budgettoekenningen of door verlaging van de in eerste instantie gemaakte productieafspraak."

Op 18 april 2003 informeert verweerder de minister over deze door hem voorgenomen afhandeling van de budgetten van de instellingen GGZ over 2003.
3. Het standpunt van verweerder

Verweerder bestrijdt het spoedeisend belang van verzoekster bij de gevraagde voorziening. De bezwaarprocedure zal naar verwachting spoedig zijn afgerond. Op 2 september 2003 zal de hoorzitting plaatsvinden. Verweerder ziet niet in waarom verzoekster de beslissing op bezwaar niet kan afwachten. Daar komt nog bij dat de reden van het verzoek louter financieel van aard is, hetgeen naar vaste rechtspraak van de voorzieningenrechter onvoldoende is om een spoedeisend belang aan te nemen. Ten slotte heeft verweerder in dit verband opgemerkt dat verzoekster geen gebruik heeft gemaakt van de in de beleidsregels voorziene mogelijkheid om aanvulling van haar budget te vragen op grond van de door haar feitelijke gerealiseerde en door het zorgkantoor gecontroleerde productie.

De toepasselijke beleidsregels zijn op de in de Wet voorgeschreven wijze tot stand gekomen. Voorts is de wijze waarop verweerder aan de beleidsregels uitvoering geeft niet onrechtmatig en

zeker niet zo onmiskenbaar onrechtmatig dat dit tot schorsing van de bestreden tariefbeschikking zou moeten leiden.

Van een groeibeperking als door verzoekster gesteld is geen sprake. Verweerder heeft een bedrag van € 3.032.920,-- gereserveerd in afwachting van het beschikbaar komen van gegevens van verzoeksters feitelijk gerealiseerde en gecontroleerde productie. Indien blijkt dat verzoekster daadwerkelijk meer zorg levert dan op basis van het aantal productieafspraken is begroot, wordt het gereserveerde bedrag, na afhandeling van het verzoek om aanvulling, als extra budget aan verzoekster ter beschikking gesteld.

Dat verzoekster het bedrag dat is gereserveerd voor productiegroei boven de 10% niet tijdelijk zou kunnen voorfinancieren acht verweerder, gelet op de ter beschikking staande cijfers omtrent haar liquiditeitspositie onwaarschijnlijk. Overigens zou, wanneer een instelling inderdaad in liquiditeitsproblemen zou verkeren, een ongeclausuleerde ophoging van het budget een onaanvaardbaar risico meebrengen. De instelling loopt dan immers de kans dat zij in de nacalculatie een aanzienlijk bedrag dat niet door gerealiseerde productie wordt gedekt moet terugbetalen. Door de behoedzame uitvoering van de beleidsregels neemt verweerder de instellingen daartegen in bescherming.

4. Het standpunt van verzoekster

Ter ondersteuning van haar verzoek heeft verzoekster - zakelijk weergegeven - het volgende betoogd.

In de Beleidsregels III-727 en III-765 is de 10% groeibeperking niet opgenomen. Een afzonderlijke goedkeuring daarvan is door de minister niet gegeven. In feite is hier sprake van een zelfstandig door verweerder gevoerd macrobeleid, dat niet strookt met het stelsel van de per 1 januari 2003 ingevoerde budgetmaximering.

Verzoekster dreigt door de beperking in de bestreden tariefbeschikking de door haar gemaakte productieafspraken niet te kunnen nakomen. Voor het realiseren van productieafspraken is het immers nodig dat personeel wordt aangetrokken, huisvesting wordt ingericht en infrastructuur wordt aangebracht. Indien de in de tariefbeschikking gemaakte reservering niet wordt opgeheven, ontbreken haar de middelen voor een adequate financiering daarvan. Haar eigen vermogen is als gevolg van het vigerende tariefsysteem beperkt. Het tarief zou overigens ook voor voorfinanciering van de voorgenomen groei de basis moeten zijn. De opstelling van verweerder verdraagt zich niet met het groeiscenario dat in de beleidsregels is voorzien.

De zorgverzekeraars kunnen verzoekster evenmin soelaas bieden, aangezien het hun niet is toegestaan hogere dan de vastgestelde tarieven te vergoeden.

Het alsnog indienen van aanvullende productieafspraken heeft ook geen zin. Een zelfde verzoek is immers reeds door haar gedaan en niet volledig gehonoreerd.

Het zorgkantoor monitort de ontwikkelingen van verzoekster nauwkeurig en beschikt over kwartaalrapportages van de productie- en wachtlijsten van verzoekster. De vage algemene zorg dat productieafspraken wel eens onvoldoende onderbouwd zouden kunnen zijn, houdt geen rekening met het uitgevoerde controlebeleid van het zorgkantoor, dat overigens heeft aangegeven te werken volgens de in het protocol gegeven aanwijzingen.

5. De beoordeling van het verzoek

5.1 Ingevolge het bepaalde bij artikel 8:81 van de Algemene wet bestuursrecht (hierna: Awb) juncto artikel 19, eerste lid, van de Wet bestuursrechtspraak bedrijfsorganisatie kan, indien tegen een besluit bij het College beroep is ingesteld, dan wel, voorafgaand aan een mogelijk beroep, bezwaar is gemaakt, op verzoek een voorlopige voorziening worden getroffen indien onverwijlde spoed, gelet op de betrokken belangen, dat vereist.

5.2 Met betrekking tot het gestelde spoedeisend belang overweegt de voorzieningenrechter als volgt. Verzoeksters financiële belang staat niet op zichzelf, maar is nauw gelieerd met de door haar gestelde noodzaak tot voorfinanciering van de voorgenomen groei door middel van het aantrekken van personeel, het inrichten van huisvesting en het aanbrengen van noodzakelijke infrastructuur. Voorshands dient, wanneer reële productieafspraken zijn gemaakt, gelet op het per 1 januari 2003 in de GGZ ingevoerde tariefsysteem waarbij de budgetmaximering is afgeschaft, voorfinanciering uit het vastgestelde budget te kunnen geschieden. Nu verzoekster onvoldoende weersproken heeft gesteld daartoe niet in staat te zijn, staat haar spoedeisend belang voldoende vast.

5.3 De voorzieningenrechter stelt vervolgens vast dat de toepasselijke beleidsregels onvoldoende grondslag bieden voor toepassing van de in deze uitspraak meermalen genoemde 10% regel. De Beleidsregel III-726 bepaalt immers alleen dat productieafspraken reëel en zonodig onderbouwd dienen te zijn en bij gebreke daarvan niet in het budget worden opgenomen. En in de Beleidsregel

III-765 is slechts bepaald dat op de productieafspraken wordt nagecalculeerd, waartoe de zorgkantoren bij de door hen uit te voeren materiële controle het productieprotocol dat als bijlage bij deze beleidsregel is gevoegd, in acht moeten nemen.

5.4 Verzoekster had dus, gelet op het bepaalde in de Beleidsregel III-726, in de gelegenheid moeten worden gesteld de realiteit van de door haar ingediende productieafspraken te motiveren. Niet is gebleken echter dat verweerder, alvorens op het tariefverzoek te beslissen, gevraagd heeft naar enige aanvullende onderbouwing van de ingediende afspraken. Verweerder heeft de bestreden tariefbeschikking genomen zonder zich nader te hebben geïnformeerd met betrekking tot de realiteit van de gemaakte budgetafspraken zoals aangeduid in beleidsregel III-726 en daarbij de door hem twee dagen eerder geformuleerde 10% regel toegepast. Voor de motivering van dat besluit heeft hij ook uitsluitend verwezen naar zijn circulaire van 15 april 2003, waarin hij de betrokkenen instellingen in algemene zin meedeelt dat een eerste inventarisatie van de ingekomen tariefverzoeken van de instellingen GGZ, hem geen duidelijk inzicht heeft gegeven in het realiteitsgehalte van de productiegroei van de onderscheiden instellingen GGZ.

5.5 De conclusie moet zijn dat verweerder niet de beleidsregels III-726 en III-765 heeft toegepast maar op structurele niet nader individueel bepaalde gronden is afgeweken van door hem vastgesteld beleid. Een bevoegdheid hiertoe kan niet aan enige regel van bestuursrecht worden ontleend. Verweerder was dan ook voorshands in de zich hier voordoende omstandigheden, gelet op het bepaalde bij artikel 4:84 van de Algemene wet bestuursrecht (Awb), gehouden tot onverkorte toepassing van de door hem vastgestelde beleidsregels.

5.6 De slotsom moet zijn dat de tariefbeschikking wegens strijd met artikel 4:84 Awb onmiskenbaar onrechtmatig is, zodat het verzoek om voorlopige voorziening voor toewijzing in aanmerking komt.

5.7 De voorzieningenrechter acht termen aanwezig voor een proceskostenveroordeling met toepassing van artikel 8:75 van de Awb.

6. De beslissing

De voorzieningenrechter:

- schorst de tariefbeschikking van verweerder van 17 april 2003, nr. 120-1740-03-3, voorzover hierin een reservering is

opgenomen ten bedrage van € 3.032.920,-- tot zes weken nadat verweerder op het bezwaar van verzoekster tegen deze

tariefbeschikking heeft beslist;

- bepaalt dat verweerder dit bedrag vooralsnog voorcalculatorisch aan verzoekster ter beschikking stelt;

-bepaalt dat verweerder het door verzoekster betaalde griffierecht ten bedrage van € 232,-- (zegge: tweehonderdtweeëndertig

euro) vergoedt;

- veroordeelt verweerder in de proceskosten van verzoekster sub 1, vastgesteld op € 644,-- (zegge: zeshonderdvierenveertig

euro).

Aldus gewezen door mr. D. Roemers, in tegenwoordigheid van mr. A. Bruining, als griffier, en uitgesproken in het openbaar op 22 juli 2003
w.g. D. Roemers w.g. A. Bruining

Samenvatting 1a

Bij tariefbeschikking van 17 april 2003, nr. 120-1740-03-2 heeft verweerder de tarieven voor verzoekster met ingang van 1 januari 2003 vastgesteld. Op 28 mei 2003 heeft verzoekster daartegen een bezwaarschrift ingediend. Op 6 juni 2003 heeft verzoekster zich tot de voorzieningenrechter van het College gewend met het verzoek een voorlopige voorziening te treffen ertoe strekkende dat de door verweerder in de bestreden beschikking gemaakte reservering wordt opgeheven en verzoeksters budget voorcalculatorisch zal worden verhoogd met € 3.032920, --.

Samenvatting 1b

Bij tariefbeschikking van 17 april 2003, nr. 120-1740-03-2 heeft verweerder de tarieven voor verzoekster met ingang van 1 januari 2003 vastgesteld. Op 6 juni 2003 heeft verzoekster zich tot de voorzieningenrechter van het college gewend met het verzoek een voorlopige voorziening te treffen ertoe strekkende dat de door verweerder in. Op 28 mei 2003 is verzoekster daartegen een bezwaarschrift ingediend. Het verzoek om voorlopige voorziening wordt afgewezen.

Evaluatie B

Uitspraak 1 (ECLI:NL:CRVB:2001:AD5001)

99/3128 AKW

U I T S P R A A K

in het geding tussen:

[A.], wonende te [B.], appellant,

en

de Sociale Verzekeringsbank, gedaagde.

I. ONTSTAAN EN LOOP VAN HET GEDING

Bij besluit van 16 oktober 1997 heeft gedaagde onder meer geweigerd aan appellant kinderbijslag toe te kennen over het vierde kwartaal van 1996 voor de kinderen [C.], geboren in 1980, [D.], geboren in 1982, [G.], geboren in 1985, en [F.], geboren in 1987.

Bij besluit op bezwaar van 9 maart 1998 is het bezwaar tegen het besluit van 16 oktober 1997 ongegrond verklaard.

De Arrondissementsrechtbank te Rotterdam heeft bij uitspraak van 10 mei 1999 het beroep tegen het besluit van 9 maart 1998 gegrond verklaard en dat besluit vernietigd voor zover daarin is gehandhaafd de beslissing dat appellant over het vierde kwartaal van 1996 geen recht heeft op kinderbijslag voor het kind [C.]. De rechtbank heeft daarbij bepaald dat gedaagde binnen zes weken nadat de uitspraak onherroepelijk is geworden een nieuw besluit dient te nemen op het bezwaarschrift met inachtneming van hetgeen de rechtbank heeft overwogen. Tevens heeft de rechtbank bepaald dat gedaagde het door appellant gestorte griffierecht aan hem vergoedt.

Appellant heeft op daartoe in het beroepschrift aangevoerde gronden hoger beroep ingesteld tegen de uitspraak van 10 mei 1999.

Gedaagde heeft een verweerschrift ingediend en daarbij tevens het besluit van 6 oktober 1999 ingezonden, waarbij appellants bezwaar tegen het besluit van 16 oktober 1997, voor zover dit betrekking heeft op het recht op kinderbijslag ten behoeve van [C.], gegrond is verklaard.

Het geding is behandeld ter zitting van de Raad, gehouden op 13 juni 2001. Appellant is daar in persoon verschenen, terwijl gedaagde, zoals was aangekondigd, zich niet heeft doen vertegenwoordigen.

II. MOTIVERING

In geschil is het antwoord op de vraag of de weigering van gedaagde om aan appellant kinderbijslag toe te kennen voor [D.], [G.] en [F.] over het vierde kwartaal van 1996 in rechte kan standhouden. De Raad gaat bij de beantwoording van die vraag uit van de volgende feiten.

Appellant is gehuwd geweest met [H.]. Op 11 januari 1996 is de echtscheiding uitgesproken en is de moeder belast met het ouderlijk gezag over de kinderen. Bij beschikking van 4 april 1996 heeft de rechtbank Rotterdam bepaald dat appellant de kinderen één dag per week bij zich mag hebben.

Op 26 januari 1997 heeft appellant kinderbijslag aangevraagd voor de kinderen. Daarbij was een afschrift gevoegd van een brief van [C.] van 24 januari 1997 aan de sector familierecht van de rechtbank Rotterdam, waarin zij onder meer meedeelt dat de kinderen vanaf 5 augustus 1996 bij appellant verblijven. Op die datum heeft hun moeder de voormalige echtelijke woning verlaten en is appellant daar weer komen wonen. [D.] heeft volgens [C.] tot november 1996 bij appellant gewoond, waarna zij bij haar moeder is gaan wonen. Op 27 mei 1997 heeft een buitendienstbeambte van gedaagde gesproken met appellant en de kinderen [F.] en [G.], die blijkens het ter zake opgemaakte rapport hebben bevestigd dat zij vanaf 5 augustus 1996 bij hun vader wonen en maar een heel enkele keer bij hun moeder op bezoek zijn geweest. Een buitendienstbeambte van gedaagdes districtskantoor Hengelo heeft op 28 mei 1997 een onderhoud gehad met de moeder van de kinderen, die sedert april 1997 in [V.] woont, waarbij zij blijkens het ter zake opgemaakte rapport heeft medegedeeld dat zij op 5 augustus 1996 met de kinderen [D.], [F.] en [G.] is verhuisd van de voormalige echtelijke woning naar de [P.]straat te [B.]. Tot de kerstvakantie verbleven de kinderen doorgaans bij haar, elk weekend en in de herfstvakantie waren de kinderen bij appellant. Het schoolgeld en de premie voor de ziektekostenverzekering zouden door haar zijn betaald. [F.] en [G.] zouden in de kerstvakantie te kennen hebben gegeven dat zij bij appellant willen blijven wonen. [D.] is bij haar moeder gebleven, terwijl [C.] vanaf augustus 1996 bij appellant woont.

Gedaagde heeft daarop het besluit van 16 oktober 1997 afgegeven en daarin overwogen dat door de tegenstrijdige verklaringen van appellant en zijn ex-echtgenote niet is vast te stellen bij wie de kinderen op 1 oktober 1996 verbleven, waardoor gedaagde genoodzaakt is om het echtscheidingsconvenant te hanteren. Daarin is bepaald dat de kinderen aan hun moeder zijn toegewezen, zodat zij op 1 oktober 1996 geacht worden tot haar huishouden te behoren. Ingaande het eerste kwartaal van 1997 heeft gedaagde wel kinderbijslag toegekend aan appellant voor [C.], [G.] en [F.].

In bezwaar heeft appellant verklaringen van drie buren overgelegd, waarin deze de lezing van appellant bevestigen. Uit onderzoek bij de gemeentelijke basisadministratie blijkt dat de kinderen van 4 juni 1996 tot 16 april 1997 waren ingeschreven op het adres [P.]straat te [B.].

Bij het thans bestreden besluit heeft gedaagde zijn standpunt gehandhaafd.

De rechtbank heeft vastgesteld dat ten aanzien van de verblijfplaats van [C.] geen tegenstrijdige verklaringen zijn afgelegd. Zij verbleef op de peildatum van het vierde kwartaal van 1996 bij appellant, zodat ten onrechte is geweigerd voor haar kinderbijslag toe te kennen aan appellant over het vierde kwartaal van 1996.

Ten aanzien van de verblijfplaats van de kinderen [D.], [G.] en [F.] op 1 oktober 1996 zijn wel tegenstrijdige verklaringen afgelegd. In zo'n geval bestaat er aanleiding doorslaggevende betekenis toe te kennen aan inhoud en strekking van ten aanzien van betrokkenen opgelegde regelingen en/of door betrokkenen duidelijk overeengekomen, objectief verifieerbare afspraken betreffende opvoeding en verblijf van een kind. Slechts indien bedoelde regelingen en/of afspraken ontbreken of te weinig uitsluitsel bieden, kunnen andere aanknopingspunten op de voorgrond treden. Nu het gezag over de kinderen aan de moeder is toegewezen en gelet op de vastgestelde omgangsregeling, heeft de rechtbank daarmee voldoende grondslag aanwezig geacht om te kunnen concluderen dat de kinderen [D.], [G.] en [F.] op 1 oktober 1996 tot het huishouden van hun moeder behoorden.

Appellant blijft bij zijn standpunt dat de kinderen ten tijde van belang bij hem verbleven.

De Raad overweegt als volgt.

Bepalend voor het antwoord op de vraag tot wiens huishouden een kind behoort is met name de feitelijke situatie. Zoals gedaagde in zijn beleidsregels heeft weergegeven behoort een kind tot het huishouden wanneer het op het adres waar het huishouden wordt gevormd het merendeel van de voor de nachtrust bestemde tijd doorbrengt. In de jurisprudentie van de Raad (o.a. gepubliceerd in RSV 1992/20) is op dat uitgangspunt in zoverre een uitzondering gemaakt dat onder omstandigheden de inhoud en strekking van de omgangsregeling bepalend is voor het antwoord op de vraag tot wiens huishouden een kind van gescheiden ouders behoort. Deze uitzondering heeft echter betrekking op de situatie waarin sprake is van co-ouderschap en daaruit voortvloeiende samenloop van het recht op kinderbijslag van beide ouders. Deze uitzondering kan niet zonder meer worden toegepast op een situatie als de onderhavige, waarin uitsluitend de ouder tot wiens huishouden het kind behoort recht heeft op kinderbijslag.

In een situatie als de onderhavige, waarin door de ouders tegenstrijdige verklaringen worden afgelegd met betrekking tot de verblijfplaats van de kinderen, rust op gedaagde een plicht om te trachten aan de hand van een onderzoek een coherent beeld van de feiten te verkrijgen. Dit onderzoek dient uitgebreider te zijn dan thans heeft plaatsgehad. Eerst wanneer ook na een zorgvuldig en gedegen onderzoek blijkt dat het niet mogelijk is vast te stellen bij wie de kinderen hebben verbleven kan aan andere gegevens, zoals een vastgestelde omgangsregeling, een zekere betekenis worden toegekend.

Uit het voorgaande volgt dat het bestreden besluit niet kan worden gedragen door de daarin vermelde motivering en voorts in strijd is met het bepaalde in artikel 3:2 van de Algemene wet bestuursrecht. Dit betekent dat de aangevallen uispraak, voor zover aangevochten, dient te worden vernietigd. Het inleidende beroep zal alsnog gegrond worden verklaard en het bestreden besluit zal, voor zover dat in stand is gelaten door de rechtbank, worden vernietigd. Gedaagde zal met inachtneming van hetgeen in deze uitspraak is vermeld een nieuw besluit op het bezwaar van appellant dienen te nemen.

Met het oog op het door gedaagde te nemen besluit merkt de Raad op dat door appellant en zijn ex-echtgenote weliswaar tegenstrijdige verklaringen zijn afgelegd, maar dat de lezing van appellant lijkt te worden bevestigd door de kinderen [C.], [G.] en [F.] en drie buren. Aan de inschrijving in het bevolkingsregister kan naar het oordeel van de Raad weinig gewicht worden toegekend, nu ook [C.] heeft ingeschreven gestaan op het adres [P.]straat te [B.], terwijl niet in geschil is dat zij daar nimmer heeft gewoond. De suggestie van appellant dat navraag zou worden gedaan bij de school van de kinderen, is door gedaagde - zonder nadere toelichting - niet opgevolgd.

Nu het hoger beroep slaagt, zal gedaagde worden veroordeeld tot vergoeding van de kosten die appellant in verband met het hoger beroep redelijkerwijs heeft moeten maken. Dit betreft de reiskosten ad f 44,58 alsmede de door appellant gevorderde verletkosten van in totaal f 600,--.

Ook dient gedaagde het door appellant in hoger beroep gestorte griffierecht ad f 170,-- aan hem te vergoeden.

Beslist wordt als volgt.

III. BESLISSING

De Centrale Raad van Beroep,

Recht doende:

Vernietigt de aangevallen uitspraak voor zover aangevochten;

Verklaart het inleidende beroep gegrond en vernietigt het bestreden besluit, voor zover dat door de rechtbank in stand is gelaten;

Bepaalt dat gedaagde met inachtneming van hetgeen in deze uitspraak is overwogen een nieuw besluit neemt op het bezwaar van appellant;

Veroordeelt gedaagde tot vergoeding van de proceskosten van appellant, begroot op f 644,58;

Bepaalt dat gedaagde het door appellant gestorte griffierecht ad f 170,-- aan hem vergoedt.

Aldus gegeven door mr. N.J. Haverkamp als voorzitter en mr. T.L. de Vries en prof.mr. W.M. Levelt-Overmars als leden, in tegenwoordigheid van mr. M.B.M. Vermeulen als griffier en uitgesproken in het openbaar op 25 juli 2001.

(get.) N.J. Haverkamp

(get.) M.B.M. Vermeulen

SS

Samenvatting 1a

Weigering kinderbijslag toe te kennen. Zoals gedaagde in zijn beleidsregels heeft weergegeven behoort een kind tot het huishouden. In de jurisprudentie van de raad (o.a. gepubliceerd in rsv 1992/20) is op dat uitgangspunt in zoverre een uitzondering gemaakt dat onder omstandigheden de inhoud. De uitzondering heeft betrekking op de situatie waarin sprake is van co-ouderschap en daaruit voortvloeiende samenloop van het recht op kinderbijslag van beide ouders. Deze uitzondering kan niet zonder meer worden toegepast op een situatie als de onderhavige, waarin uitsluitend de ouder tot wiens huishouden het kind behoort.

Samenvatting 1b

Gescheiden ouders, omgangsregeling, verblijfplaats kinderen

F.3 Questions

Vragen

Evaluatie A

1a) Relevantie (0-10):

Leesbaarheid (0-10):

1b) Relevantie (0-10):

Leesbaarheid (0-10):

Onderbouwing:

Evaluatie B

1) Gebruik je liever 1a of 1b?:

1a: Relevantie (0-10):

Leesbaarheid (0-10):

Onderbouwing:

General information			Performance					Human evaluation			Technical details				Documentation & Opportunities	
Technique	**Authors**		**Dataset**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Training**	**Method**	**Results**		**Framework**	**Libraries**	**Core value**		**Documentation**	**Opportunities**
Fast Abstractive Rewriting	Chen & Bansal (2018)		Non-anonymized CNN/Daily Mail	40.88%	17.80%	38.54%	19.71 hours (See et al. (2017) had 78 hours). Used Nvidia Tesla K40 GPU for training.	Measured relevance, readability and abstractiveness of the model. Used Amazon MTurk to review 100 samples. Summaries were randomly shuffled and reviewers rated a summary as A is better, B is better or both are equally good/bad.	Had a higher abstractiveness (novel n-gram count) compared to See et al. (2017). Both relevance and readability were 10% and 5% better than See et al. (2017),		Torch (0.4.0)	Python 3.6. Word2Vec embeddings form gensim package with 30K vocabulary. CUDA 9	The model first selects important sentences and then rewrites them abstractively. Reinforcement Learning techniques from Paulus are used. Speeds up training time drastically compared to other models		Pre-trained model is available, as well as a large set of guidelines for training your own models. Pipeline of the training and summary generation is well defined	Use a new and improved version of the Stanford CoreNLP Tokenizer. Use a new version of CUDA to improve GPU performance. Re-write to Torch 1.0 library. Increase the vocabulary size (might be doable if our GPU is better). Tune hyperparameters
Controllable Abstractive Summarization	Fan et al. (2017)		Non-anonymized CNN/Daily Mail	40.38%	17.44%	37.15%	-	Used Amazon Mturk to review 500 samples from the test set, evaluated by 5 raters. The raters were presented with the first 400 words of each news article and asked to select the summarization output they preferred.	Human raters prefer their model about 59% of the time, compared to reference summaries. For the model of See et al. (2017), this was only 41%.		Torch (0.2.0**)	Python 3.6+. BPE embeddings with a 30K vocabulary. CUDA 8	Options are available to determine the desired length of the summaries, what entities the model needs to focus on and style of the summary (e.g. a particular news source)		Their is no documentation explaining how to implement their model. The Github only offers general guidelines for the sequence-to-sequence learning toolkit for translation.	Increase vocabulary size. Tune hyperparameters. Rewrite to Torch 1.0 library and use a new version of CUD
Reinforcement Learning	Paulus et al., 2017)		Anonymized CNN/Daily Mail	39.87%	15.82%	36.90%	-	Used Amazon Mturk to review 100 samples from the test set. Two scores from 1 to 10 are then assigned to each summary, one for relevance and one for readability. Each summary is rated by 5 different human evaluators.	Their Reinforcement Learning model produced the highest ROUGE, but low readability. By adding a Mixed Learning Objective, ROUGE decreased, but readability and relevance went up by a large margin.		TensorFlow (1.10.1)	Python 2.7. GloVe embeddings, 150K vocab. CUDA 9	Implements Reinforcement Learning techniques to look back to previously generated inputs and make sure the model does not repeat itself		Another author, Keneshloo et al. (2018), implemented the model of Paulus et al. (2017) in their paper. He provides guidelines for replicating their experiment as well as options for parameters, but the documentation is not very expansive	Rewrite to newest version of CUDA and Tensorflow 2.0. Tune hyperparameters

Figure 17: Criteria and results for determining the best abstractive summarization for this research