




Multi-language transfer learning for low-resource legal case summarization

Gianluca Moro¹ · Nicola Piscaglia² · Luca Ragazzi¹  · Paolo Italiani¹

Accepted: 5 June 2023
© The Author(s) 2023

Abstract

Analyzing and evaluating legal case reports are labor-intensive tasks for judges and lawyers, who usually base their decisions on report abstracts, legal principles, and commonsense reasoning. Thus, summarizing legal documents is time-consuming and requires excellent human expertise. Moreover, public legal corpora of specific languages are almost unavailable. This paper proposes a transfer learning approach with extractive and abstractive techniques to cope with the lack of labeled legal summarization datasets, namely a low-resource scenario. In particular, we conducted extensive multi- and cross-language experiments. The proposed work outperforms the state-of-the-art results of extractive summarization on the Australian Legal Case Reports dataset and sets a new baseline for abstractive summarization. Finally, syntactic and semantic metrics assessments have been carried out to evaluate the accuracy and the factual consistency of the machine-generated legal summaries.

Keywords Legal case reports · Extractive summarization · Abstractive summarization · Transfer learning · Multi-language · NLP

✉ Gianluca Moro
gianluca.moro@unibo.it

✉ Luca Ragazzi
l.ragazzi@unibo.it
Nicola Piscaglia
nicola.piscaglia@bbs.unibo.it
Paolo Italiani
paolo.italiani@unibo.it

¹ Department of Computer Science and Engineering - DISI, University of Bologna, Cesena Campus, Via dell'Università 50, 47522 Cesena, Italy

² Bologna Business School, University of Bologna, Via degli Scalini, 18, 40136 Bologna, Italy

1 Introduction

Automatic text summarization is undoubtedly one of the most valuable deep learning applications for natural language processing (NLP), especially when document analysis is time-consuming for humans. Reading and evaluating legal case reports is labor-intensive for judges and lawyers, who usually base their choices on report abstracts, legal principles, and commonsense reasoning. Legal report abstracts are poorly available, and the text summarization task requires law experts and a long time to be performed. Furthermore, few datasets are publicly ready for text summarization in the legal domain, mainly if we need corpora written in a specific language because legislation is drafted in its relative nation's language. Therefore, the challenge of this work is to build an automatic summarization system of legal reports to speed up human productivity and overcome the dearth of available legal corpora.

Two main approaches have been proposed in the literature concerning the text summarization task: extractive and abstractive. Extractive summarization aims to select the most salient sentences in a text, obtaining a summary with exact sentences from the original document. Instead, the abstractive approach summarizes an input text by rephrasing it, also using new words that may not be present in the source text.

In this work, both extractive and abstractive summarization techniques have been tackled by proposing a transfer learning approach that can be used to cope with the lack of labeled legal summarization datasets (i.e., legal corpora without human-generated abstracts), which is a typical low-resource scenario concerning the available data. Our method allows generating abstractive summaries by just starting from tagged catchphrases within legal reports (Fig. 1). The catchphrases are meant to present the essential legal concepts of a case. To this end, we first select and extract the relevant sentences in the text, if not already tagged, by using a lightweight neural model composed of CNN and GRU layers. Then, we pass the extracted sentences to GPT-2 (Radford et al. 2019) as the reference summary. We chose GPT-2 because it is a generative decoder-only transformer-based model that is more efficient than usual sequence-to-sequence summarization models based on encoder and decoder layers that double the memory space and training time.

Our approach can be considered a general solution to overcome the absence of abstractive reference summaries written by human legal experts. Indeed, only catchphrases tags are required, which are much less time-consuming to create than human-written abstracts because they can be obtained in multiple ways, e.g., by applying an unsupervised extractive algorithm (e.g., TextRank (Mihalcea and Tarau 2004)) or by manually tagging the sentences of a few documents and afterward fine-tuning a pre-trained model.

We experiment with the Australian Legal Case Reports dataset, evaluating our model's effectiveness in summarizing legal case reports in several languages (i.e., Australian, Italian, German, Spanish, Danish, French) whose translations are yielded automatically with the Google Translate API as in prior works (Feng et al. 2022). We also test the cross-language performance by directly summarizing the

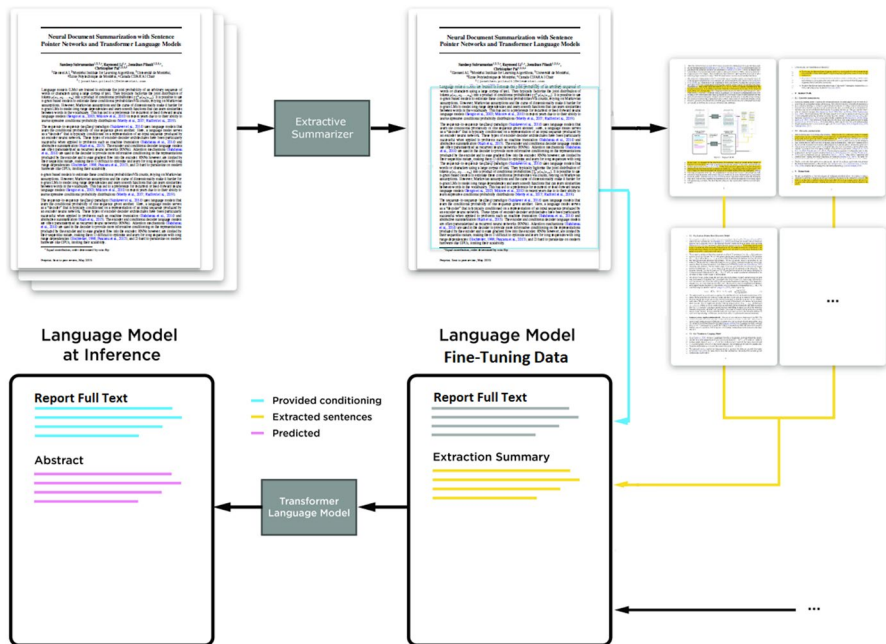


Fig. 1 The overview of our solution for the abstractive summarization of a legal case report. First, our extractive summarizer (CNN+GRU) retrieves relevant sentences from the document. Next, these sentences are provided along with the whole legal case report to be arranged in the following order: (i) Report Full Text, (ii) Extracted Sentences. Then, a transformer language model (GPT-2 small) is trained on legal case reports organized in such a format. During inference, the full text of each new legal case report to summarize is given to the language model as context to generate a summary

English-written texts in all the benchmarked languages. We finally assess the factual consistency (Kryscinski et al. 2020) of the generated abstractive summaries.

The paper is organized as follows. Section 2 analyzes the literature about text summarization and deeps into the related works on the Australian Legal Case Reports dataset. Section 3 presents our transfer learning approach and models employed for the summarization tasks. Section 4 shows the multi-language experiments with extractive and abstractive techniques. Lastly, Sect. 5 sums up the work with final thoughts.

2 Related work

In order to achieve state-of-the-art (SOTA) results in automatic text summarization, many advancements have been made in neural network architectures (Cho et al. 2014; Sutskever et al. 2014; Bahdanau et al. 2015; Vinyals et al. 2015; Vaswani et al. 2017; Moro and Valgimigli 2021; Moro et al. 2022), pretraining and transfer learning (Domeniconi et al. 2015, Domeniconi et al. 2016, McCann et al. 2017; Peters et al. 2018; Devlin et al. 2019), and the availability of large-scale supervised

datasets (Sandhaus 2008; Nallapati et al. 2016; Grusky et al. 2018; Narayan et al. 2018; Sharma et al. 2019), which allowed deep learning-based approaches (Domeniconi et al. 2017) to dominate the field, also for biomedical tasks (Frisoni et al. 2021; Frisoni et al. 2022; Frisoni et al. 2023) and multi-modal settings (Moro and Salvatori 2022; Moro et al. 2023), and overcome the need of low-resource summarization techniques (Moro and Ragazzi 2022; Moro et al. 2023a, b; Moro and Ragazzi 2023). SOTA solutions leverage attention layers (Liu 2019; Liu and Lapata 2019; Zhang et al. 2019), copying mechanisms (See et al. 2017; Cohan et al. 2018), and multi-objective training strategies (Guo et al. 2018; Pasunuru and Bansal 2018), including reinforcement learning (RL) techniques (Kryscinski et al. 2018; Dong et al. 2018; Wu and Hu 2018).

SOTA Extractive summarization includes BERT-based models, such as BertSum (Liu and Lapata 2019), BERT + RL (Bae et al. 2019), and PnBert (Zhong et al. 2019). Conversely, abstractive summarization includes transformer-based models for sequence-to-sequence learning, such as BART (Lewis et al. 2020), PEGASUS (Zhang et al. 2020), T5 (Raffel et al. 2020), and ProphetNet (Qi et al. 2020). Further, in the new research focused on building efficient transformers with linear complexity, the SOTA models in long document summarization are Longformer Encoder-Decoder (Beltagy et al. 2020), BigBird-Pegasus (Zaheer et al. 2020), and Hepos (Huang et al. 2021).

The summarization task on the Australian Legal Case Reports dataset has been first tackled by (Galgani and Hoffmann 2010). They presented a new knowledge-based approach towards legal citation classification and created an extensive training and test corpus from court decision reports in Australia. Their later work (Galgani et al. 2012) presents the challenges and possibilities for the automatic generation of catchphrases for legal documents. The authors developed a corpus of human-generated legal catchphrases, which lets them compute statistics useful for automatic catchphrase extraction. Afterward, (Galgani et al. 2012a) presented their approach to assigning categories and generating catchphrases for legal case reports. They describe their knowledge acquisition framework, which lets them quickly build classification rules, using a small number of features to assign general labels to cases. They show how the resulting knowledge base outperforms machine learning models, which use both the designed features or a traditional bag of words representation. In the same year, (Galgani et al. 2012b) described a hybrid approach in which several different summarization techniques are combined in a rule-based system using manual knowledge acquisition. Here, human intuition, supported by data, specifies attributes and algorithms and the contexts where these are best used. Lastly, (Galgani et al. 2012c) presented an approach towards using both incoming and outgoing citation information to generate catchphrases automatically for legal case reports. Specifically, they created a corpus of cases, catchphrases, and citations and performed a ROUGE-based evaluation (Lin et al. 2004), which showed the superiority of their citation-based methods over full-text-only methods.

Mandal et al. (2017) proposed an unsupervised approach for extracting and ranking catchphrases from the same court case documents by focusing on noun phrases. They compared the proposed approach with several unsupervised and

supervised baselines, showing that the proposed methodology achieves statistically significantly better performance than all the baselines.

In the latest published work on Australian legal cases, Tran et al. (2018) presented a method of automatic catchphrase extraction. They utilized deep neural networks to construct a scoring model of their extraction system and achieved comparable performance without using citation information.

Other recent works focused on benchmarking different extractive and abstractive approaches (Shukla et al. 2022) without considering catchphrases extraction or experimenting with it on different datasets (Koboyatshwene et al. 2017; Bhargava et al. 2017; Kayalvizhi and Thenmozhi 2020).

Our work focused on the Australian Legal Case Reports dataset for two main reasons. First, several studies have already been performed on such a dataset. Second, it is suitable for simulating the lack of human-crafted abstracts, unlike the BillSum dataset (Kornilova and Eidelman 2019), for example. Further, this work has been partially inspired by a recent approach proposed by Pilault et al. (2020), which combines extractive summarization with abstractive one to increase the performance of their model. Nevertheless, since we do not have abstracts as reference summaries, we fine-tune a pre-trained transformer-based model using extractive summaries as reference ones to overcome that absence.

3 Method

In this work, we apply extractive and abstractive summarization techniques to propose a transfer learning approach to generate abstractive summaries starting from tagged catchphrases. In particular, BERT (Devlin et al. 2019) (base, multilingual cased) and a deep neural network composed of CNN and GRU layers have been used for the extraction phase, whereas the abstractive one has been performed with the GPT-2 transformer-based model.

3.1 Extractive summarization

In order to generate contextualized word embeddings, we apply the BERT-Base-Multilingual-Cased pre-trained model for two main reasons: (i) BERT has achieved the SOTA in various NLP tasks, and (ii) the multilingual model allows us to overcome the absence of multi-lingual legal case datasets in the cross-language experiments. We obtained the sentence embeddings by carrying out the mean of the word embeddings related to words within the same sentence.

For the binary classification task, our model (Fig. 2) comprises:

1. One 1D CNN layer (LeCun et al. 1999), with a subsequent MaxPooling1D operation, with a kernel of size 1 and filters of size 1024.

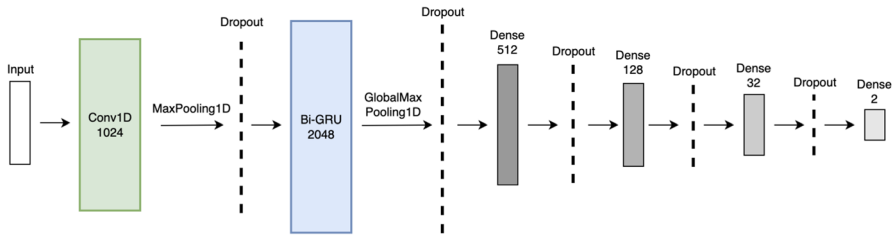


Fig. 2 The model architecture used in the extractive summarization experiments

2. One bidirectional GRU layer (Cho et al. 2014) with a GlobalMaxPooling1D operation.
3. Four fully connected layers of decreasing dimensionality.

All the main layers are interleaved with Dropout levels. Technically, Conv1D, MaxPooling1D, and Dropout layers have only been applied when the mean of word embeddings has been used as the sentence embedding building method for performance reasons.

The word/sentence embeddings have been yielded using the Flair NLP library (Akbi et al. 2019).¹ This framework lets us choose among different embedding methods and generate tensors of words/sentences from the relative string.

We train the extractive summarization model to minimize the categorical cross-entropy loss, as follows:

$$\mathcal{L}_{es} = - \sum_{i=1}^{i=N} y_i \cdot \log(\hat{y}_i) \quad (1)$$

where N is the number of samples and y_i and \hat{y}_i are the gold and predicted labels, respectively.

3.2 Abstractive summarization

GPT-2 has been used to generate legal abstractive summaries in the abstractive summarization scenario. To this end, two main steps are required:

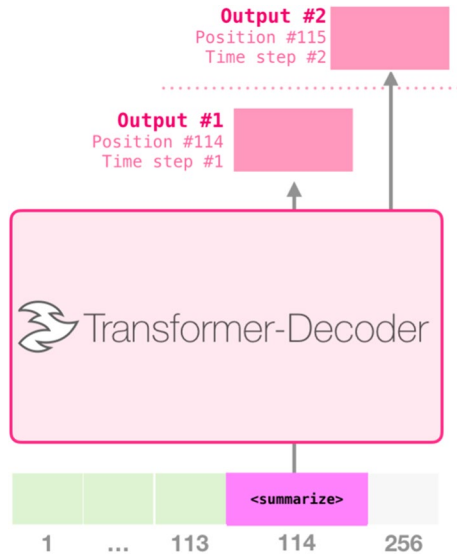
1. Fine-tuning data (Fig. 3): each instance of our GPT-2 fine-tuning data is sequentially composed of the **input text**, a `<summarize>` **tag**, and the **input text summary**. In our case, the input text will be the original text of each legal case report, whereas the summary will be the set of extracted sentences labeled as relevant for each report.
2. Inference phase (Fig. 4): each test data instance is composed of the **original text** of a new legal report followed by the `<summarize>` **tag**.

¹ <https://github.com/flairNLP/flair>.

Article #1 tokens	<summarize>	Article #1 Summary
Article #2 tokens	<summarize>	Article #2 Summary
		padding
Article #3 tokens	<summarize>	Article #3 Summary

Fig. 3 The structure of each instance of fine-tuning data used in the abstractive summarization with the GPT-2 language model. Each data is composed of: (i) the legal case report full text, (ii) a <summarize> tag, and (iii) the report summary obtained via the extractive summarization process

Fig. 4 The inference phase in the abstractive summarization process performed via GPT-2. For each summary to produce, the full text of the legal case report is concatenated to the <summarize> tag. The GPT-2 model iterates a specified number of steps producing one token at a time. The report summary will be the tokens generated after the <summarize> tag



The embedding process is integrated into the GPT-2 architecture, where sequences of words are transformed into numeric vectors by the tokenizer.

We train the abstractive summarization model using the standard cross-entropy loss, which requires the model to predict the next token y_i of the target \mathcal{Y} given \mathcal{X} and the previous target tokens $y_{1:i-1}$, as follows:

$$\mathcal{L}_{as} = - \sum_{i=1}^{|\mathcal{Y}|} \log p(y_i | y_{1:i-1}, \mathcal{X}) \tag{2}$$

where p is the predicted probability over the vocabulary.

4 Experiments

This section first introduces the dataset, experimental setup, and evaluation metrics. Then, we deep into the in- and cross-domain experiments for extractive and abstractive approaches.

4.1 Dataset

The dataset used in our experiments is the Australian Legal Case Reports, and it represents a textual corpus of around 4000 legal cases for automatic summarization and citation analysis (Galgani et al. 2012c). The dataset contains Australian legal cases from the Federal Court of Australia (FCA) from 2006 to 2009, downloaded from AustLII.² For each document, the authors collected catchphrases, citation sentences, citation catchphrases, and citation classes that indicate the type of treatment given to the cases cited by the present case.

The dataset is structured in three directories:

- **fulltext**: it contains the full text and the catchphrases of all the cases from the FCA. Each document (`<case>`) contains:
 - `<name>`: the name of the case.
 - `<AustLII>`: the link to the page from where the document was taken.
 - `<catchphrases>`: a list of `<catchphrase>` elements.
 - `<sentences>`: a list of `<sentence>` elements.
- **citations_summ**: it contains citations element for each case with the following fields:
 - `<name>`: the name of the case.
 - `<AustLII>`: the link to the page from where the document was taken.
 - `<citphrases>`: a list of `<citphrase>` elements that are catchphrases from a case which is cited or cite the current one. The attributes are `id`, `type` (cited or citing), and `from` (the case from where the catchphrase is taken).
 - `<citances>`: a list of `<citance>` elements that are sentences from a later case that mention the current case. They also have the `from` attributes.
 - `<legistitles>`: a list of `<title>` elements that are titles of a piece of legislation cited by the current case.
- **citations_class**: it contains for each case a list of labeled citations with the following fields:
 - `<name>`: the name of the case.
 - `<AustLII>`: the link to the page from where the document was taken.
 - `<citations>`: a list of `<citation>` elements. They contains several attributes, such as the `<class>` of the citation as indicated on the docu-

² <http://www.austlii.edu.au/>.

ment (considered, followed, cited, applied, notfollowed, referred to, etc.), the name of the case which is cited (<tocase>), the link to the document of the case which is cited (<AustLII>), and the <text> paragraphs in the cited case where the current case is mentioned.

No missing values have been found in `fulltext` and `citation_class` directory files, whereas some values are missing in `citation_summ` documents.

XML files contain many HTML entity characters. These latter ones made XML parsing invalid since “&” characters would indicate entities of XML types and not HTML ones, so they had to be removed.

The data used to perform the analysis have been selected from the `fulltext` directory. For each legal case (i.e., an XML file), <name>, <catchphrase>, and <sentence> have been used. HTML special entities have been removed to parse the text correctly. To this end, we replaced HTML entity characters with the corresponding textual representation. Some legal case reports have been truncated as they are not encoded as UTF-8 strings.

In order to create the target variable (i.e., the feature representing the class) of our extractive summarization experiments, a label for each legal sentence is needed, indicating whether a sentence should be included in a case report summary. Thus, the class variable will be binary. Since this information is not directly specified in the metadata, it has been generated using the following annotation process: for each sentence of a legal case, it is checked whether at least one of the catchphrases for that legal case is included in the current legal sentence examined. If this condition is true, then the label of that sentence will be `True`, else it will be `False`.

The legal sentences of each case report have been added to a common dataframe. So, each instance of this latter structure will represent a phrase. The instances in the dataframe have been balanced by class (represented by `is_catchphrase` attribute). Afterward, data instances were shuffled by groups of phrases from the same legal case report to avoid the situation where the model will classify phrases from legal case reports already seen during training time. Table 1 shows statistics of all datasets (i.e., the Australian Legal Case Reports translated into all the evaluation languages), reporting the number of words and sentences in source and target texts, source-target compression ratio (the number of source words divided by the number of target words), and the % of relevant sentences containing catchphrases.

4.2 Experimental setup

Similar to previous contributions (Zhang et al. 2020), all the analyses have been performed both on a dataset of sizes 100 and 1000, simulating real-world scenarios characterized by the dearth of data. Precisely, one table has been generated for (i) each summarization technique adopted and (ii) each language to which the dataset has been translated. As commonly applied in realistic organizations because of the high data labeling cost, all extractive and abstractive models have been trained using 70% of the dataset sampled in each experiment and tested on the remaining samples, similar to Bajaj et al. (2021).

Table 1 Statistics of the legal case reports translated into all the evaluation languages

	Source		Target		% Catchphrases	Compression ratio
	# Words	# Sentences	# Words	# Sentences		
100 legal case reports						
Australian	6597	251	590	15	6.69	38.62
Italian	6081	232	606	16	7.01	27.37
German	5522	216	450	11	6.58	38.42
Spanish	5346	195	459	11	6.55	33.98
Danish	4326	174	353	9	6.60	33.98
French	5602	206	462	11	6.59	38.94
1000 legal case reports						
Australian	7365	280	615	16	6.22	36.42
Italian	6869	267	579	15	6.57	37.74
German	6591	257	541	14	6.44	36.53
Spanish	7094	257	587	14	6.44	35.94
Danish	7777	311	639	17	6.44	37.21
French	6898	252	572	14	6.44	35.67

Regarding the multi-lingual experiments, we translate each legal case report using the Google Translate API in the following languages: Italian, German, Spanish, Danish, and French.

4.2.1 Implementation details

We fine-tuned the models based on the Keras implementations for tensors computation, setting the seed to 7 for reproducibility. We trained the extractive summarization model for 100 epochs with a batch size of 50 and a learning rate of $1e-4$, employing Adam as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and a weight decay of $1e-6$. We used the L1 and L2 regularization penalties of 0 and 0.001, respectively, and setting the dropout to 0.1. Regarding abstractive summarization, we train the model for 4 epochs with a batch size of 4 and a learning rate of $5e-5$, using Adam as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and a weight decay of $1e-6$. For decoding, we utilized top-p nucleus sampling with top_p and temperature set to 0.9 and 1, respectively.

4.3 Evaluation metrics

The extractive experiments have been tested by evaluating the F1 and ROUGE scores (Lin et al. 2004). The F1 score is calculated between the sentences classified as relevant and the gold ones containing the catchphrases. F1 metric has been computed to consider both recall (i.e., the ratio of relevant sentences retrieved by the model out of all relevant sentences in the dataset) and precision metrics (i.e.,

Table 2 The F1 scores obtained in the evaluation of several word embedding methods in the extractive summarization scenario

Examples #	Word		Context	
	100	1000	100	1000
Models				
Word2Vec	67.10	72.69	85.62	87.83
GloVe	64.93	73.31	81.24	87.82
ELMo	74.83	75.89	79.69	88.35
BERT-Base-Multilingual	70.41	74.22	75.46	90.29
BERT-Large	70.92	78.33	83.92	85.22

“Context” refers to using token embeddings’ sequences as input to create contextual embeddings. The best scores are bolded

the percentage of salient sentences retrieved out of all the sentences in the produced summary). Further, we choose the F1 metric because both the training and test set are not perfectly class-balanced since—after the class-balancing operation—data is grouped by case report groups and shuffled. By doing so, we also simulate a production environment where an entirely new legal case report is passed as input to our classifier. Conversely, the ROUGE scores are calculated between the concatenation of the classified sentences and the concatenation of the gold-relevant ones.

The evaluation metrics used for the abstractive summarization task are ROUGE and FactCC (Kryscinski et al. 2020). The latter is used to evaluate the factual consistency of the summaries w.r.t. the related report’s original texts. In particular, the F1 score and balanced accuracy metrics have been calculated. Technically, we used the authors’ data generation scripts to generate positive and negative examples from a JSONL file. Negative examples are created by applying some syntactic transformations to the original texts.

4.4 Extractive summarization

In order to evaluate the numerous embedding methods among those proposed in the literature, several extractive summarization experiments were carried out (Table 2). BERT-Large is the best performer in creating embeddings of single words without surrounding context, with a 78.33% F1. The intuition behind the gap between BERT-Base-Multilingual and BERT-Large models is that the latter has been trained only on English text and has a larger dimension. However, we chose BERT-Base-Multilingual because it has a similar performance and it can be used as a baseline to compare the results in multi-lingual settings. On the other hand, models performed better where sequences as taken into account for creating contextualized word embeddings because orders are not lost with such a sentence representation, leading to higher F1 scores. In this case, BERT-Base-Multilingual is the best performer. Thus, this latter has been chosen since it has the highest performance and can be used as a baseline to compare the results of the experiments involving more than one language.

4.4.1 In-domain single-language experiments

The detailed results of the in-domain extractive summarization tasks on the original Australian Legal Case Reports dataset are shown in Table 3. As expected, the ROUGE scores obtained in the experiments with 1000 legal case reports are the highest. These ROUGE scores outperform, on average, the ones produced by the latest work on the same dataset, which used a neural network for the catchphrase extraction task (Tran et al. 2018).

4.4.2 In-domain multi-language experiments

Table 4 shows the results of the extractive summarization performed on 100 and 1000 legal case reports translated into Italian, German, Spanish, Danish, and French. Regarding 100 samples, the experiment with the Spanish dataset achieved the best results in almost all metrics except for the F1 score, which was obtained with the Danish dataset. Considering 1000 samples, the best performances were found in the German translation scenario except for the F1 score, which was achieved with the Italian dataset.

Figures 5 and 6 sum up the results of the extractive experiments. Table 5 and Table 6 compare the multi-lingual results of other extractive summarization approaches, revealing the better performance of our solution. In detail, we compare with MemSum (Gu et al. 2022), BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and DistilBERT (Sanh et al. 2019). For all languages, we use the

Table 3 The ROUGE (Precision, Recall, F1) and F1 scores obtained in the extractive summarization experiments using the original English-written reports and the two different strategy as the building method of sentence embeddings. The best scores are in bold

Examples #	100			1000		
Mean of word embeddings						
F1 Score	70.41			74.22		
ROUGE-1	P: 42.01	R: 47.82	F1: 43.18	P: 47.99	R: 54.35	F1: 49.51
ROUGE-2	P: 26.42	R: 30.04	F1: 26.94	P: 32.17	R: 35.57	F1: 32.91
ROUGE-3	P: 23.24	R: 26.57	F1: 23.64	P: 28.79	R: 31.60	F1: 29.39
ROUGE-4	P: 22.17	R: 25.44	F1: 22.53	P: 27.73	R: 30.44	F1: 28.28
ROUGE-L	P: 44.62	R: 50.09	F1: 45.89	P: 49.79	R: 55.33	F1: 51.26
vROUGE-w1.2	P: 31.02	R: 17.85	F1: 21.26	P: 34.91	R: 19.90	F1: 24.20
Token embedding sequences						
F1 Score	75.46			90.29		
ROUGE-1	P: 60.27	R: 59.70	F1: 59.98	P: 57.36	R: 70.45	F1: 60.54
ROUGE-2	P: 47.45	R: 51.55	F1: 48.53	P: 47.72	R: 56.62	F1: 50.07
ROUGE-3	P: 45.17	R: 49.70	F1: 46.96	P: 45.40	R: 53.57	F1: 47.58
ROUGE-4	P: 44.69	R: 49.35	F1: 46.56	P: 44.60	R: 52.64	F1: 46.73
ROUGE-L	P: 62.41	R: 60.97	F1: 61.68	P: 59.90	R: 71.72	F1: 63.09
ROUGE-w1.2	P: 45.82	R: 22.10	F1: 28.33	P: 41.73	R: 25.46	F1: 29.86

Table 4 The ROUGE-F1 and F1 scores obtained in the extractive summarization experiments using 100 and 1000 Australian Legal Case Reports, including the original ones (written in English) and those translated into several languages (Italian, German, Spanish, Danish, French)

	F1	R-1	R-2	R-3	R-4	R-L	R-w1.2
100 legal case reports							
Australian	70.41	43.18	26.94	23.64	22.53	45.89	21.26
Italian	65.20	61.29	50.43	48.78	47.87	63.91	32.36
German	66.64	47.10	33.79	31.43	30.40	50.00	23.33
Spanish	69.68	77.73	69.60	67.48	66.28	79.86	40.78
Danish	69.78	60.69	50.15	48.00	46.95	63.44	32.16
French	67.97	59.85	46.82	44.11	42.82	61.36	29.34
1000 legal case reports							
Australian	74.22	49.51	32.91	29.39	28.28	51.26	24.20
Italian	70.90	64.74	55.68	53.83	52.79	67.69	34.68
German	69.62	67.70	59.85	58.07	56.96	70.68	36.28
Spanish	69.63	66.88	53.71	50.79	49.59	69.06	33.64
Danish	70.01	62.55	52.08	49.77	48.38	65.61	32.01
French	69.80	55.73	40.64	37.45	36.14	57.73	26.77

The mean of word embeddings has been used as the building method for sentence embeddings. The best scores are bolded

corresponding multi-lingual model checkpoint (except for MemSum because it is available only for English and RoBERTa that does not include Danish versions).

4.4.3 Cross-language experiments

Table 7 shows the results of the extractive summarization task in the cross-domain scenario. The fine-tuning technique has been used as the transfer learning approach. The following experiments have been conducted:

- Regarding the experiments with 100 samples, (1) we tested the extractive model trained on 70 English reports directly on 30 cases of a different language, (2) we fine-tuned the extractive model (already trained on 70 English documents) with 70 reports of different languages and then tested it on 30 cases of a different language.
- About the experiments with 1000 samples, (1) we tested the model trained on 700 English samples directly on 30 cases of a different language, (2) we fine-tuned the model (trained on 700 English reports) with 70 documents of different languages and tested it on 30 cases of a different language.

Our models show a good generalization capability across different language domains. The application of the fine-tuning technique has boosted the considered evaluation metrics, showing how transfer learning can be used to overcome the lack of a labeled dataset in a specific language, as in the legal domain.

Table 5 The comparison with extractive summarization models on the multi-language experiment with 100 labeled samples. Best scores are bolded

100 samples	F1	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-w1.2
Australian							
MemSum	20.47	32.53	15.93	12.90	12.30	26.67	11.03
BERT	28.05	35.43	23.33	21.15	20.67	32.48	14.84
RoBERTa	30.17	35.78	24.07	21.93	21.40	32.72	15.19
DistilBERT	32.19	34.72	23.97	21.76	21.21	31.85	15.11
Our	70.41	43.18	26.94	23.64	22.53	45.89	21.26
Italian							
BERT	43.38	39.83	31.37	30.25	29.84	38.16	18.92
RoBERTa	44.55	38.80	31.04	29.91	29.43	37.77	18.72
DistilBERT	41.73	39.83	31.00	30.10	29.80	38.49	19.06
Our	65.20	61.29	50.43	48.78	47.87	63.91	32.36
German							
BERT	33.74	28.01	20.70	19.42	18.85	27.18	12.83
RoBERTa	18.96	18.67	15.32	14.84	14.68	18.22	9.21
DistilBERT	27.82	27.87	20.37	19.05	18.45	27.00	12.70
Our	66.64	47.10	33.79	31.43	30.40	50.00	23.33
Spanish							
BERT	24.61	30.46	19.13	17.14	16.33	27.54	12.04
RoBERTa	32.64	33.57	22.16	20.00	19.23	30.00	13.62
DistilBERT	28.47	30.87	19.98	17.84	17.07	27.85	12.33
Our	69.68	77.73	69.60	67.48	66.28	79.86	40.78
Danish							
BERT	28.58	28.63	19.57	18.23	17.62	25.99	11.81
DistilBERT	27.52	24.40	15.81	14.57	14.04	22.85	9.74
Our	69.78	60.69	50.15	48.00	46.95	63.44	32.16
French							
BERT	32.52	29.10	18.28	16.16	15.43	25.86	10.99
RoBERTa	25.50	32.96	21.62	19.47	18.69	28.86	13.05
DistilBERT	31.26	33.37	23.17	21.04	20.28	30.04	13.90
Our	67.97	59.85	46.82	44.11	42.82	61.36	29.34

4.5 Abstractive summarization

4.5.1 In-domain single-language experiments

The results of the abstractive summarization tasks performed on the original Australian Legal Case Reports dataset are shown in Table 8. As expected, in the scenario with 1000 samples, the results are much higher for each metric than the ones obtained in

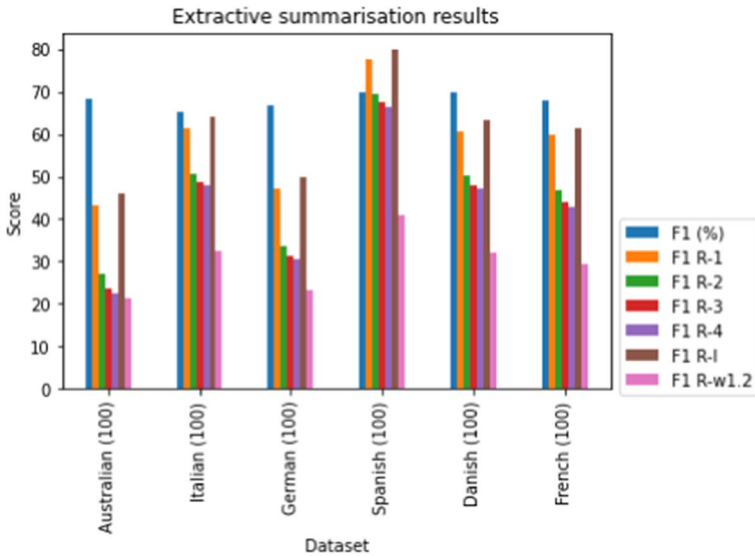


Fig. 5 The ROUGE-F1 and F1% scores obtained in the extractive summarization experiments using 100 Australian legal case reports translated into multiple languages. The mean of word embeddings has been used as the building method for sentence embeddings

the experiment with 100 reports. The results are similar to those obtained in the latest catchphrase extraction work proposed in the literature on the same dataset (Tran et al. 2018). This gives us the intuition that our abstractive model can produce abstracts with a certain degree of lexical and syntactic correctness.

4.5.2 In-domain multi-language experiments

Table 9 shows the results of the abstractive summarization performed on 100 and 1000 translated reports of the Australian Legal Case Reports dataset. Regarding 100 samples, the best ROUGE-3 score is achieved with the French dataset, whereas we obtained the best results from the experiment with the Spanish dataset. Considering 1000 documents, the best scores are achieved in the Spanish translation scenario for all metrics except for ROUGE-3 and ROUGE-4, where the experiment with the French dataset got the first place.

Figures 7 and 8 sum up the results of the abstractive experiments.

4.5.3 FactCC assessment

Table 10 shows the results of the FactCC model evaluation on 60 and 1800 Australian legal case reports scenarios, respectively. The legal case report number is doubled since their approach provides transformations, which creates another dataset containing negative examples. As expected, the model fine-tuned with more

Table 6 The comparison with extractive summarization models on the multi-language experiment with 1000 labeled samples. Best scores are bolded

1000 samples	F1	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-w1.2
Australian							
MemSum	28.58	34.94	23.17	20.65	19.76	30.79	13.68
BERT	46.50	42.74	34.39	32.77	32.32	40.53	20.66
RoBERTa	46.02	39.83	34.21	33.22	32.93	38.65	20.36
DistilBERT	46.92	41.60	33.61	32.13	31.70	39.63	20.21
Our	74.22	49.51	32.91	29.39	28.28	51.26	24.20
Italian							
BERT	44.85	35.70	29.27	28.38	28.02	34.20	17.55
RoBERTa	42.67	34.46	28.03	27.18	26.81	33.11	16.94
DistilBERT	44.75	35.63	29.22	28.39	28.04	34.21	17.51
Our	70.90	64.74	55.68	53.83	52.79	67.69	34.68
German							
BERT	43.80	33.01	27.54	26.80	26.52	32.02	16.48
RoBERTa	23.14	26.90	21.08	20.21	19.86	25.82	12.82
DistilBERT	44.88	33.65	28.01	27.27	26.99	32.53	16.72
Our	69.62	67.70	59.85	58.07	56.96	70.68	36.28
Spanish							
BERT	47.40	37.51	30.64	29.45	29.07	36.15	18.25
RoBERTa	47.68	39.82	31.64	30.27	29.80	37.79	19.07
DistilBERT	47.02	39.13	30.97	29.48	28.99	36.82	18.50
Our	69.63	66.88	53.71	50.79	49.59	69.06	33.64
Danish							
BERT	28.15	20.89	18.33	18.05	17.90	20.60	10.75
DistilBERT	39.45	29.22	23.89	23.27	23.00	27.99	14.23
Our	70.01	62.55	52.08	49.77	48.38	65.61	32.01
French							
BERT	43.03	34.81	28.15	27.11	26.76	32.81	16.55
RoBERTa	35.69	33.06	25.92	24.63	24.13	30.83	15.14
DistilBERT	45.56	38.53	30.89	29.67	29.24	36.29	18.19
Our	69.80	55.73	40.64	37.45	36.14	57.73	26.77

examples (2100) achieves the best results among the two experiments, with a much higher balanced accuracy and F1 score. This makes us think the model needs many training data to replicate or overcome the results obtained in the original paper (Kryscinski et al. 2020). As the FactCC model has performed well on legal cases, it has been used as the metric for evaluating our generated abstractive summaries.

We aim to evaluate our abstractive summaries by running the previously trained FactCC model (with 2100 legal training reports). This latter has been used to

classify 100 and 1000 abstractive machine-generated summaries as `CORRECT` or `INCORRECT`. If a summary is evaluated as `CORRECT`, we have reasonable assurance that it is fluently and coherently written. The training technique is BERT-based-uncased fine-tuning for 8 epochs with the default parameters of the FactCC model. Table 11 shows the ratio of report summaries classified as `CORRECT` out of all the evaluated summaries. In evaluating 650 abstractive summaries, nearly 50% of them are classified as consistent.

Although the limits of the evaluation method, we believe this result shows how GPT-2 can produce abstracts with a certain degree of consistency even in the legal domain. In addition, this evaluation represents a baseline, which may be improved with future summarization technique enhancements.

4.6 Evaluation

4.6.1 Extractive summarization

In order to compare our results with the SOTA, we searched all the works which used the Australian Legal Case Reports dataset. The latest one we found is Tran et al. (2018) (Table 12), and it has been used as the baseline for the extractive summarization tasks because it has analogies with our work:

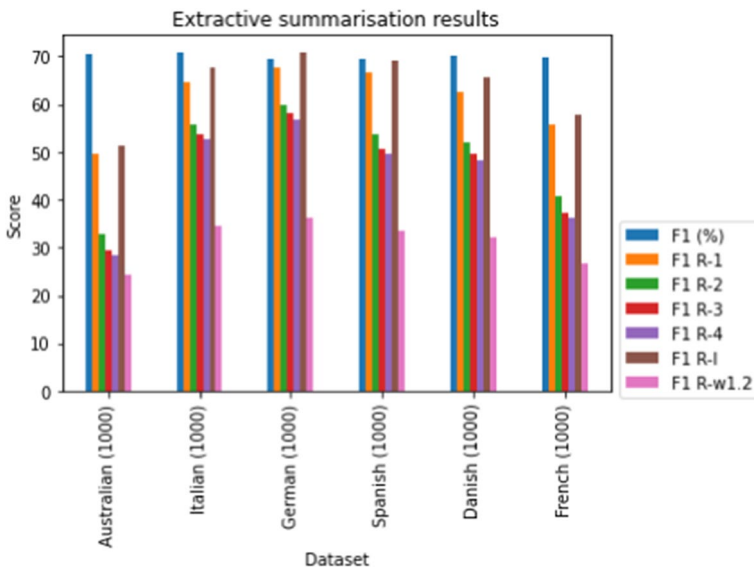


Fig. 6 The ROUGE-F1 and F1% scores obtained in the extractive summarization experiments using 1000 Australian legal case reports translated into multiple languages. The mean of word embeddings has been used as the building method for sentence embeddings

Table 7 The ROUGE (F1) and F1 scores of the cross-language experiments, where our extractive model trained on English reports has been applied in two experiments: (i) it has been tested without fine-tuning on different languages (EN \rightarrow LN); (ii) it has been fine-tuned using the translated legal case reports and then tested (EN, LN \rightarrow LN)

LN	F1	R-1	R-2	R-3	R-4	R-L	R-w1.2
70 EN \rightarrow 30 LN							
Italian	64.43	57.49	44.37	42.11	40.85	60.93	29.67
German	68.63	53.30	42.93	41.06	40.12	56.35	28.02
Spanish	64.88	51.58	36.65	33.64	32.51	54.24	25.54
Danish	64.81	51.78	40.00	37.68	36.40	54.96	26.25
French	71.28	63.61	53.10	50.71	49.46	66.24	33.53
70 EN, 70 LN \rightarrow 30 LN							
Italian	83.07	64.61	53.12	50.93	49.90	66.47	34.25
German	79.25	50.45	37.72	35.63	34.78	53.40	26.33
Spanish	85.32	65.71	54.65	52.91	52.13	68.07	34.33
Danish	78.33	54.78	45.58	43.86	42.85	58.09	28.15
French	86.99	67.09	57.64	55.58	54.61	69.29	35.32
700 EN \rightarrow 30 LN							
Italian	56.65	53.33	39.22	36.66	35.76	56.36	28.12
German	57.38	31.67	18.04	16.34	15.69	34.80	15.01
Spanish	64.52	45.70	28.28	25.25	24.39	47.68	21.38
Danish	60.06	32.33	18.57	16.85	15.93	34.85	14.90
French	62.82	48.33	34.16	31.86	30.91	50.83	23.71
700 EN, 70 LN \rightarrow 30 LN							
Italian	83.27	63.05	52.00	49.55	48.54	65.56	33.15
German	90.32	51.44	39.79	37.99	37.07	54.61	27.24
Spanish	86.29	64.58	53.09	50.91	49.87	66.99	33.27
Danish	85.70	45.27	31.45	29.66	28.63	47.81	22.13
French	80.00	62.08	51.72	49.72	48.65	64.61	32.16

The mean of word embeddings has been used as the building method for sentence embeddings

- It used the same data.
- It did not involve citation data in the training process.
- It only used sentences and words from catchphrases as the target data.

By comparing our results, it can be stated that our model achieved excellent performance in syntactic terms. In particular, ROUGE-1 and ROUGE-W1.2 scores obtained in our experiments are much higher. The primary motivation could be using BERT as the word/sentence embedding builder system. As explained in the first chapter, choosing a good contextualized embedding model is crucial for better performance. Another reason could be using a more expressive classification model: we used LSTM networks combined with CNN ones, whereas, in their work, only CNNs have been applied. Extractive summarization of translated reports achieves better results for almost all metrics than the English scenario. In particular, the best scores have been obtained by the experiment with the Spanish dataset for the

Table 8 The ROUGE (Precision, Recall, F1) obtained in the abstractive summarization using the original English-written reports. The best scores are in bold

Examples #	100			1000		
ROUGE-1	P: 26.55	R: 21.51	F1: 22.84	P: 28.93	R: 26.18	F1: 27.18
ROUGE-2	P: 3.30	R: 2.76	F1: 2.88	P: 4.86	R: 4.43	F1: 4.59
ROUGE-3	P: 0.17	R: 0.14	F1: 0.15	P: 0.99	R: 0.92	F1: 0.94
ROUGE-4	P: 0.00	R: 0.00	F1: 0.00	P: 0.42	R: 0.41	F1: 0.42
ROUGE-L	P: 28.10	R: 23.29	F1: 24.70	P: 30.37	R: 27.93	F1: 28.87
ROUGE-w1.2	P: 14.74	R: 5.71	F1: 7.93	P: 15.68	R: 7.06	F1: 9.64

Table 9 The ROUGE-F1 scores obtained in the abstractive summarization experiments using 100 and 1000 Australian Legal Case Reports, including the original ones (written in English) and those translated into several languages (Italian, German, Spanish, Danish, French). The best scores are bolded

	R-1	R-2	R-3	R-4	R-L	R-w1.2
100 legal case reports						
Australian	22.84	2.88	0.15	0.00	24.70	7.93
Italian	22.00	1.19	0.10	0.00	24.25	8.06
German	21.83	2.51	0.41	0.03	24.12	8.02
Spanish	33.47	4.93	0.61	0.19	36.31	11.97
Danish	22.89	1.67	0.31	0.03	25.85	8.45
French	30.18	4.15	0.78	0.18	31.42	10.02
1000 legal case reports						
Australian	27.18	4.59	0.94	0.42	28.87	9.64
Italian	25.11	2.75	0.56	0.15	27.59	9.25
German	25.31	4.02	1.01	0.26	28.06	9.47
Spanish	35.70	6.57	1.56	0.53	37.52	12.38
Danish	25.39	2.83	0.73	0.23	27.90	9.15
French	32.44	6.14	1.83	0.76	33.46	10.97

ROUGE metrics and the Danish dataset for the F1 score. This gives us the intuition that the model benefits from using the BERT multilingual model as the embedding builder. This latter lets us generate expressive contextualized word embeddings and allow us to work with different languages. In Table 13, we showcase a representative qualitative instance for each of the languages analyzed thus far. The efficacy of our solution in extracting sentences that include catchphrases is readily apparent.

4.6.2 Abstractive summarization

Since no similar works for the abstractive summarization on the same dataset have been found, extractive summarization results have been used as the baseline. As expected, ROUGE scores of the abstractive summarization tasks are much lower than the extractive summarization ones. Indeed, the ROUGE score is a mere lexical measure. In the extractive scenario, if the model succeeds in classifying one sentence as relevant or not, then all the words of that sentence represent an overlapping

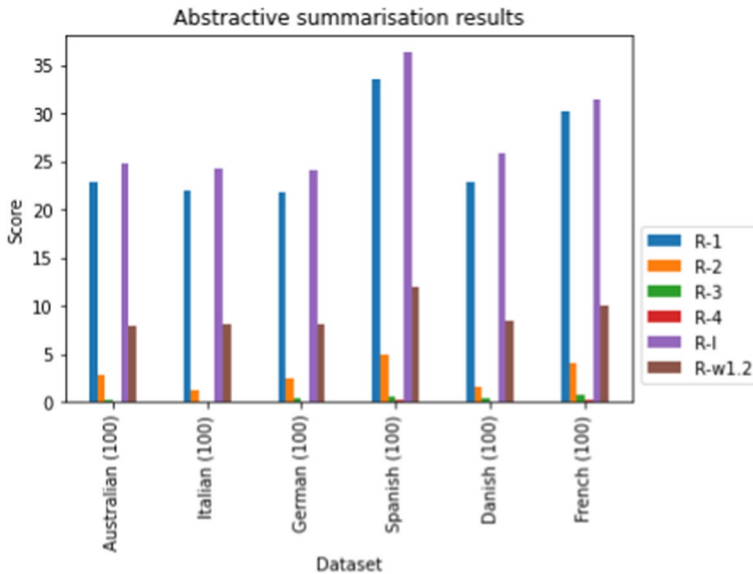


Fig. 7 The ROUGE-F1 and F1% Scores obtained in the abstractive summarization experiments using 100 Australian legal case reports translated into multiple languages

between the generated and the reference summary. This does not apply to the abstractive summarization task by definition because here, the goal is to produce a new summary, also using words that do not exist in the input text to summarize. However, the ROUGE scores of the abstractive summarization experiments are similar to Tran et al. (2018), so we have the intuition that our abstractive model has been able to generate speeches, which are inherent to the input text even if consistency and fluidity of speech are to verify yet. In order to do that, the FactCC model has been applied, and it turned out that about 46% of 300 machine-generated reports have been classified as CORRECT. Even though our fine-tuned FactCC model has a 77% F1 score (i.e., it is affected by errors), this fact gives us the intuition that our abstractive summaries have a certain degree of fluency and coherency w.r.t. their related legal report original texts, which are not reflected in the ROUGE rating. Abstractive summarization of translated reports achieves similar results to the English scenario and even better for Spanish and French languages. This gives us the intuition that the model keeps working well in languages other than English and could be applied to many use cases.

5 Conclusion

In this work, we tackled the automatic summarization of Australian legal case reports by presenting extractive and abstractive techniques. The abstractive solution can be considered a general approach to generating summaries despite lacking human-crafted references. Our method only requires catchphrase tags that can be

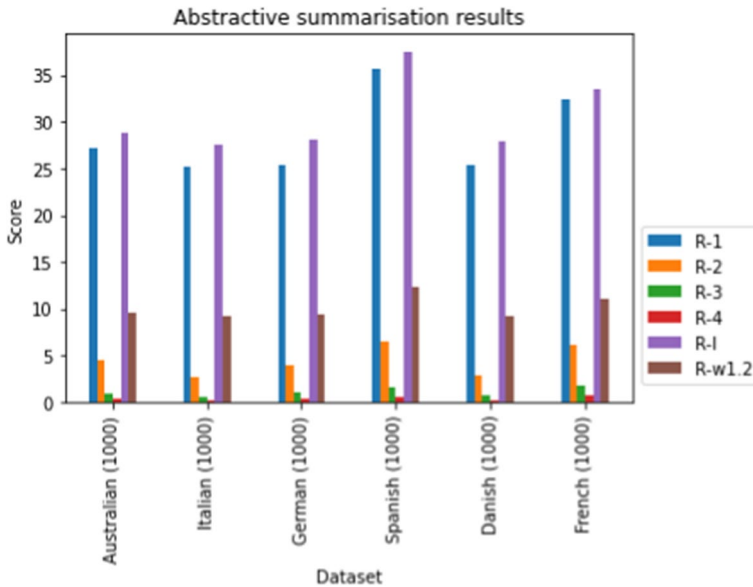


Fig. 8 The ROUGE-F1 and F1% Scores obtained in the abstractive summarization experiments using 1000 Australian legal case reports translated into multiple languages

obtained in several ways: (i) by applying an unsupervised extractive summarization algorithm or (ii) by manually tagging the sentences of a few documents and afterward fine-tuning a pre-trained model.

We showed that our extractive summarization results overtake the ones produced by the latest work on the same dataset in the literature (Tran et al. 2018). Instead, our abstractive summarization results led to similar ROUGE scores to theirs. In addition, we proved the speech consistency of our abstractive summaries using the FactCC model. Precisely, even though our fine-tuned FactCC model can make inference errors on test data (77% F1), the results suggest that our abstractive summaries have a certain degree of fluency and coherency w.r.t. their related legal sources, which are not reflected in the ROUGE rating.

Moreover, a translation task has been achieved to train our models and evaluate their ability to understand and summarize texts in several languages other than English. It turned out that the summarization of translated reports achieves better results than the English report scenario for some other languages. Especially, Spanish and French seem to perform generally better in the abstractive summarization case. In contrast, the summarization of German reports achieves the best results in the extractive summarization scenario with 1000 reports. Hence, our models summarize several languages effectively and could be applied to other legal case reports. Such experiments are supported by Google Translate API and the BERT multilingual embedding model (only used in the extractive summarization scenario). Finally, it turned out that our models can also generalize in a cross-language scenario, summarizing English reports directly in all the different languages.

Table 10 The balanced accuracy and F1 scores of the FactCC model assessment after the fine-tuning using the Australian Legal Case Reports dataset (sampling 100 and 3000 reports, respectively)

Examples #	60	1800
Balanced Accuracy (%)	50.00	76.91
F1 Score (%)	66.15	77.50

Table 11 The percentage of inferred abstractive summaries classified as CORRECT (i.e., consistent with the related original full text) by our fine-tuned FactCC model in 3 experiments with different reports sampling (30, 300, and 650 machine-generated summaries via abstraction)

Predictions #	30	300	650
CORRECT ratio (%)	41.67	46.70	49.92

Table 12 The ROUGE-1 and ROUGE-W-1.2 (Precision, Recall, F1) scores from “Automatic Catchphrase Extraction from Legal Case Documents via Scoring using Deep Neural Networks” by Tran et al. (2018)

ROUGE-1	P: 23.11	R: 30.84	F1: 22.95
ROUGE-w1.2	P: 14.50	R: 13.63	F1: 11.75

The main challenges will be improving the quality of the machine-generated abstractive summaries and their evaluation. Some automatic evaluations like (Kryscinski et al. 2020) have been proposed in the literature and represent an improvement to the previous solutions to this problem, even though they still have limitations.

Future works will be related to replicating the cross-language experiments using token embedding sequences as input instead of applying the mean of word embeddings to build sentence embeddings since the first method has led to higher results in the in-domain experiments with original Australian legal reports. Furthermore, the methods proposed in this work could be expanded by adding more advanced data techniques to FactCC to improve abstractive summaries evaluation and repeat the experiments using other SOTA large language models (e.g., ChatGPT) to improve the abstraction quality. Finally, as presented for communication network (Lodi et al. 2010; Moro and Monti 2012; Cerroni et al. 2013; Cerroni et al. 2015), propagating knowledge refinements (Domeniconi et al. 2014), also with entity relationships acquisition (Frisoni et al. 2020; Frisoni and Moro 2021) and event extraction (Frisoni et al. 2021), could be key when modeling complex long legal documents.

Table 13 Qualitative examples of extractive summarization in the multilingual setting. Catchphrases are highlighted in italics

Language	Input Document	Extractive summary
Australian	The circumstances revealed in this application for <i>preliminary discovery</i> suggest... Especially is this so where the person concerned is an Australian citizen and the information.	The circumstances revealed in this application for preliminary discovery suggest.
Italian	Il sig. Gormly sostiene che la <i>negazione della giustizia naturale</i> "contagia" l'intera seconda decisione del Tribunale.... Il sig. Johnson, che si presenta per il ministro, sostiene che la base alternativa per il.	Il sig. Gormly sostiene che la negazione della giustizia naturale "contagia" l'intera.
German	Der vorliegende Fall wird von zwei Behörden in Bezug auf verstorbene Personen und <i>Antidiskriminierungsgesetze</i> sorgfältig eingegrenzt.... In Stephenson gegen Human Rights and Equal Opportunity Commission (1996).	Der vorliegende Fall wird von zwei Behörden in Bezug auf verstorbene Personen und Antidiskriminierungsgesetze sorgfältig eingegrenzt.
Spanish	...En Inglaterra, la <i>prueba pericial</i> se restringe a la que se requiere razonablemente para resolver los procedimientos. La licencia se concede o deniega como parte de la gestión del caso.	...En Inglaterra, la prueba pericial se restringe a la que se requiere razonablemente para resolver los procedimientos.
Danish	Dette er en <i>ansøgning om udeblivelsesdom</i> i henhold til ordre 35A i Federal Court Rules... .. Den blev serveret fredag den 3. februar 2006, hvilket er mindre end de tre fridage.	Dette er en ansøgning om udeblivelsesdom i henhold til ordre 35A i Federal Court Rules.
French	...Les membres du groupe de <i>demande d'indemnisation</i> sont presque exclusivement Yankunyvtajajara ou Pijanjajajara... ..Les requérants admettent que l'un des cinq requérants restants, Mantajajara Wilson, ne satisfait pas.	Les membres du groupe de demande d'indemnisation sont presque exclusivement Yankunyvtajajara ou Pijanjajajara.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R (2019) FLAIR: an easy-to-use framework for state-of-the-art NLP. In: Ammar W, Louis A, Mostafazadeh N (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Demonstrations. Association for Computational Linguistics, pp 54–59. <https://doi.org/10.18653/v1/n19-4010>
- Bae S, Kim T, Kim J, Lee S (2019) Summary level training of sentence rewriting for abstractive summarization. [arXiv:1909.08752](https://arxiv.org/abs/1909.08752)
- Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference track proceedings. <http://arxiv.org/abs/1409.0473>
- Bajaj A, Dangati P, Krishna K, Kumar PA, Uppaal R, Windsor B, Brenner E, Dotterrer D, Das R, McCallum (2021) A Long document summarization in a low resource setting using pretrained language models. In: Kabbara J, Lin H, Paullada A, Vamvas J (eds) Proceedings of the ACL-IJCNLP 2021 student research workshop. ACL, pp 71–80. <https://doi.org/10.18653/v1/2021.acl-srw.7>
- Beltagy I, Peters ME, Cohan A (2020) Longformer: the long-document transformer. [arXiv:2004.05150](https://arxiv.org/abs/2004.05150)
- Bhargava R, Nigwekar S, Sharma Y (2017) Catchphrase extraction from legal documents using LSTM networks. In: Majumder P, Mitra M, Mehta P, Sankhavara J (eds) Working Notes of FIRE 2017—proceedings of forum for information retrieval evaluation, Bangalore, India, December 8–10, 2017. CEUR Workshop , vol 2036, pp 72–73. <http://ceur-ws.org/Vol-2036/T3-3.pdf>
- Cerroni W, Moro G, Pirini T, Ramilli M (2013) Peer-to-peer data mining classifiers for decentralized detection of network attacks. In Proceedings of the Twenty-Fourth Australasian Database Conference 137:101–107
- Cerroni W, Moro G, Pasolini R, Ramilli M (2015) Decentralized detection of network attacks through P2P data clustering of SNMP data. *Computers & Security* 52:1–16
- Cho K, van Merriënboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, a meeting of SIGDAT, a Special Interest Group of The ACL, pp 1724–1734. <https://doi.org/10.3115/v1/d14-1179>
- Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N (2018) A discourse-aware attention model for abstractive summarization of long documents. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, vol 2 (Short Papers) Association for computational linguistics, pp 615–621. <https://doi.org/10.18653/v1/n18-2097>
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human

- language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- Dong Y, Shen Y, Crawford E, van Hoof H, Cheung JCK (2018) Banditsum: extractive summarization as a contextual bandit. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, October 31–November 4, 2018. Association for Computational Linguistics, pp 3739–3748. <https://doi.org/10.18653/v1/d18-1409>
- Domeniconi G, Maseroli M, Moro G, Pinoli P (2016) Cross-organism learning method to discover new gene functionalities. In *Computer methods and programs in biomedicine* 126:20–34
- Domeniconi G, Moro G, Pagliarani A, Pasolini, R (2015) Markov chain based method for in-domain and cross-domain sentiment classification. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) IEEE 1:127–137
- Domeniconi G, Moro G, Pagliarani A, Pasolini R (2017) On Deep Learning in Cross-Domain Sentiment Classification. In *KDIR* pp 50–60
- Domeniconi, G, Moro, G, Pasolini, R, Sartori, C (2014) In Cross-domain Text Classification through Iterative Refining of Target Categories Representations. In *KDIR* pp 31–42
- Feng X, Feng X, Qin B (2022) MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In: Proceedings of the second DialDoc workshop on document-grounded dialogue and conversational question answering. Association for Computational Linguistics, Dublin, Ireland, pp 1–12. <https://doi.org/10.18653/v1/2022.dialdoc-1.1>
- Frisoni G, Moro G (2021) Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In *Data Management Technologies and Applications: 9th International Conference, DATA 2020, Virtual Event, July 7–9, 2020, Revised Selected Papers 9* (pp. 293–318). Springer International Publishing.
- Frisoni G, Moro G, Carbonaro A (2020) Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining. In *DATA* pp 121–132
- Frisoni G, Moro G, Carbonaro A (2021) A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access* 9:160721–160757
- Frisoni G, Italiani P, Salvatori S, Moro G (2023) Cogito ergo summ: abstractive summarization of biomedical papers via semantic parsing graphs and consistency rewards. In: Proceedings of the AAAI Conference on Artificial Intelligence 37(11) 12781–12789
- Frisoni G, Mizutani M, Moro G, Valgimigli L (2022) Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In: Proceedings of the 2022 conference on empirical methods in natural language processing pp 5770–5793
- 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. Association for Computational Linguistics, pp 115–123
- Galgani F, Compton P, Hoffmann AG (2012) Citation based summarisation of legal texts. In: Anthony P, Ishizuka M, Lukose D (eds) *PRICAI 2012: trends in artificial intelligence—proceedings of 12th Pacific Rim international conference on artificial intelligence*, Kuching, Malaysia, September 3–7, 2012. Lecture Notes in Computer Science, vol. 7458. Springer, pp 40–52. https://doi.org/10.1007/978-3-642-32695-0_6
- Galgani F, Compton P, Hoffmann AG (2012) Knowledge acquisition for categorization of legal case reports. In: Richards D, Kang BH (eds) *Knowledge management and acquisition for intelligent systems—12th Pacific Rim knowledge acquisition workshop, PKAW 2012*, Kuching, Malaysia, September 5–6, 2012. Lecture Notes in Computer Science, vol 7457. Springer, pp 118–132. https://doi.org/10.1007/978-3-642-32541-0_10
- Galgani F, Compton P, Hoffmann AG (2012) Towards automatic generation of catchphrases for legal case reports. In: Gelbukh AF (ed) *Computational linguistics and intelligent text processing—13th international conference, CICLing 2012*, New Delhi, India, March 11–17, 2012, Part II. Lecture notes in computer science, vol 7182. Springer, pp 414–425. https://doi.org/10.1007/978-3-642-28601-8_35
- Galgani F, Hoffmann AG (2010) LEXA: towards automatic legal citation classification. In: Li J (ed) *Proceedings of AI 2010: advances in artificial intelligence—23rd Australasian joint conference*, Adelaide, Australia, December 7–10, 2010. Lecture Notes in Computer Science, vol 6464. Springer, pp 445–454. https://doi.org/10.1007/978-3-642-17432-2_45

- Grusky M, Naaman M, Artzi Y (2018) Newsroom: a dataset of 1.3 million summaries with diverse extractive strategies. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers). Association for Computational Linguistics, pp 708–719. <https://doi.org/10.18653/v1/n18-1065>
- Gu N, Ash E, Hahnloser R (2022) MemSum: extractive summarization of long documents using multi-step episodic Markov decision processes. In: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 6507–6522. <https://doi.org/10.18653/v1/2022.acl-long.450>
- Guo H, Pasunuru R, Bansal M (2018) Soft layer-specific multi-task summarization with entailment and question generation. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers. Association for Computational Linguistics, pp 687–697. <https://doi.org/10.18653/v1/P18-1064>
- Huang L, Cao S, Parulian N.N, Ji H, Wang L (2021) Efficient attentions for long document summarization. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tür D, Beltagy I, Bethard S, Cotterell R, Chakraborty T, Zhou Y (eds) Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2021, Online, June 6–11, 2021. Association for Computational Linguistics, pp 1419–1436. <https://doi.org/10.18653/v1/2021.naacl-main.112>
- Kayalvizhi S, Thenmozhi D (2020) Deep learning approach for extracting catch phrases from legal documents. In: Neural networks for natural language processing. IGI Global, pp 143–158
- Koboyatshwene T, Lefoane M, Narasimhan L (2017) Machine learning approaches for catchphrase extraction in legal documents. In: Majumder P, Mitra M, Mehta P, Sankhavara J (eds) Working Notes of FIRE 2017–forum for information retrieval evaluation, Bangalore, India, December 8–10, 2017. CEUR workshop proceedings, vol 2036, pp 95–98. <http://ceur-ws.org/Vol-2036/T3-11.pdf>
- Kornilova A, Eidelman V (2019) Billsun: a corpus for automatic summarization of US legislation. [arXiv:1910.00523](https://arxiv.org/abs/1910.00523)
- Krystinski W, McCann B, Xiong C, Socher R (2020) Evaluating the factual consistency of abstractive text summarization. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, November 16–20, 2020. Association for Computational Linguistics, pp 9332–9346. <https://doi.org/10.18653/v1/2020.emnlp-main.750>
- Krystinski W, Paulus R, Xiong C, Socher R (2018) Improving abstraction in text summarization. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, October 31–November 4, 2018. Association for Computational Linguistics, pp 1808–1817. <https://doi.org/10.18653/v1/d18-1207>
- LeCun Y, Haffner P, Bottou L, Bengio Y (1999) Object recognition with gradient-based learning. In: Forsyth DA, Mundy JL, Gesù VD, Cipolla R (eds) Shape, contour and grouping in computer vision. Lecture Notes in Computer Science, vol 1681. Springer. https://doi.org/10.1007/3-540-46805-6_19
- Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics, pp 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. Association for Computational Linguistics, Barcelona, Spain, pp 74–81. <https://www.aclweb.org/anthology/W04-1013>
- Liu Y (2019) Fine-tune BERT for extractive summarization. [arXiv:1903.10318](https://arxiv.org/abs/1903.10318)
- Liu Y, Lapata M (2019) Text summarization with pretrained encoders. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019. Association for Computational Linguistics, pp 3728–3738. <https://doi.org/10.18653/v1/D19-1387>
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach. [arXiv:1907.11692](https://arxiv.org/abs/1907.11692)

- Lodi S, Moro G, Sartori C (2010) Distributed data clustering in multi-dimensional peer-to-peer networks. In Proceedings of the Twenty-First Australasian Conference on Database Technologies 104:171–178
- Mandal A, Ghosh K, Pal A, Ghosh S (2017) Automatic catchphrase identification from legal court case documents. In: Lim E, Winslett M, Sanderson M, Fu AW, Sun J, Culpepper JS, Lo E, Ho JC, Donato D, Agrawal R, Zheng Y, Castillo C, Sun A, Tseng VS, Li C (eds) Proceedings of the 2017 ACM on conference on information and knowledge management, CIKM 2017, Singapore, November 06–10, 2017. ACM, pp 3728–3738. <https://doi.org/10.1145/3132847.3133102>
- McCann B, Bradbury J, Xiong C, Socher R (2017) Learned in translation: Contextualized word vectors. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, Long Beach, CA, USA, pp 6294–6305. <https://proceedings.neurips.cc/paper/2017/hash/20c86a628232a67e7bd46f76fba7ce12-Abstract.html>
- Mihalcea R, Tarau P (2004) Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, EMNLP 2004, a meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25–26 July 2004, Barcelona, Spain. ACL, pp 404–411. <https://www.aclweb.org/anthology/W04-3252/>
- Moro G, Monti G (2012) W-Grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. In Journal of Network and Computer Applications 35(4):1218–1234
- Moro G, Ragazzi L (2023) Align-then-abstract representation learning for low-resource summarization. *Neurocomputing* 545:126356
- Moro G, Ragazzi L, Valgimigli L, Frisoni G, Sartori C, Marfia G (2023) Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*. <https://doi.org/10.3390/s23073542>
- Moro G, Ragazzi L (2022) Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: AAAI 2022, virtual event, February 22 - March 1, 2022. AAAI Press, pp 11085–11093
- Moro G, Ragazzi L, Valgimigli L (2023) Carburacy: summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In: Proceedings of the AAAI Conference on Artificial Intelligence 2023, Washington, DC, USA, February 7–14, 2023. AAAI Press, 37(12):14417–14425
- Moro G, Ragazzi L, Valgimigli L, Freddi D (2022) Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. *ACL (Volume 1: Long Papers)*. ACL, Dublin, pp 180–189. <https://doi.org/10.18653/v1/2022.acl-long.15>
- Moro G, Salvatori S (2022) Deep Vision-Language Model for Efficient Multi-modal Similarity Search in Fashion Retrieval. *International Conference on Similarity Search and Applications*. Springer International Publishing, Cham, pp 40–53
- Moro G, Salvatori S, Frisoni G (2023) Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval. In: *Neurocomputing*, 538, 126196.
- Moro G, Valgimigli L (2021) Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature. In: *Sensors* 21(19):6430. <https://doi.org/10.3390/s21196430>
- Nallapati R, Zhou B, dos Santos CN, Gülçehre Ç, Xiang B (2016) Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Goldberg Y, Riezler S (eds.) Proceedings of the 20th SIGNLL conference on computational natural language learning, CoNLL 2016, Berlin, Germany, August 11–12, 2016. ACL, pp 280–290. <https://doi.org/10.18653/v1/k16-1028>
- Narayan S, Cohen SB, Lapata M (2018) Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, October 31–November 4, 2018. Association for Computational Linguistics, pp 1797–1807. <https://doi.org/10.18653/v1/d18-1206>
- Pasunuru R, Bansal M (2018) Multi-reward reinforced summarization with saliency and entailment. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, pp 646–653. <https://doi.org/10.18653/v1/n18-2102>
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language

- technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, vol 1 (Long Papers). Association for Computational Linguistics, pp 2227–2237. <https://doi.org/10.18653/v1/n18-1202>
- Pilault J, Li R, Subramanian S, Pal C (2020) On extractive and abstractive neural document summarization with transformer language models. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, Online, November 16–20. Association for Computational Linguistics, pp 9308–9319. <https://doi.org/10.18653/v1/2020.emnlp-main.748>
- Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, Zhang R, Zhou M (2020) Prophetnet: predicting future n-gram for sequence-to-sequence pre-training. In: Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing: findings, EMNLP 2020, Online Event, 16–20 November 2020. Findings of ACL, vol EMNLP 2020. Association for Computational Linguistics, pp 2401–2410. <https://doi.org/10.18653/v1/2020.findings-emnlp.217>
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8)
- Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 21(140):1–67
- Sandhaus E (2008) The New York times annotated corpus. Linguistic Data Consortium, Philadelphia 6(12):26752
- Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. [arXiv:1910.01108](https://arxiv.org/abs/1910.01108)
- See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: Barzilay R, Kan M (eds) Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, vol 1: Long Papers. Association for Computational Linguistics, pp 1073–1083. <https://doi.org/10.18653/v1/P17-1099>
- Sharma E, Li C, Wang L (2019) BIGPATENT: a large-scale dataset for abstractive and coherent summarization. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, vol 1: Long Papers. Association for Computational Linguistics, pp 2204–2213. <https://doi.org/10.18653/v1/p19-1212>
- Shukla A, Bhattacharya P, Poddar S, Mukherjee R, Ghosh K, Goyal P, Ghosh S (2022) Legal case document summarization: extractive and abstractive methods and their evaluation. In: He Y, Ji H, Liu Y, Li S, Chang C, Poria S, Lin C, Buntine WL, Liakata M, Yan H, Yan Z, Ruder S, Wan X, Arana-Catania M, Wei Z, Huang H, Wu J, Day M, Liu P, Xu R (eds) Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing, AACL/IJCNLP 2022–vol 1: Long Papers, Online Only, November 20–23, 2022. Association for Computational Linguistics, pp 1048–1064. <https://aclanthology.org/2022.aacl-main.77>
- Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp 3104–3112. <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>
- Tran VD, Nguyen ML, Satoh K (2018) Automatic catchphrase extraction from legal case documents via scoring using deep neural networks. [arXiv:1809.05219](https://arxiv.org/abs/1809.05219)
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp 2692–2700. <https://proceedings.neurips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html>
- Wu Y, Hu B (2018) Learning to extract coherent summary via deep reinforcement learning. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second aaAI conference on artificial intelligence,

- (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, pp 5602–5609. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16838>
- Zaheer M, Guruganesh G, Dubey K.A, Ainslie J, Alberti C, Ontañón S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A (2020) Big bird: Transformers for longer sequences. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual. <https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html>
- Zhang H, Cai J, Xu J, Wang J (2019) Pretraining-based natural language generation for text summarization. In: Bansal M, Villavicencio A (eds) Proceedings of the 23rd conference on computational natural language learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019. Association for Computational Linguistics, pp 789–797. <https://doi.org/10.18653/v1/K19-1074>
- Zhang J, Zhao Y, Saleh M, Liu PJ (2020) PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In: Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, Virtual Event. Proceedings of machine learning research, vol 119. PMLR, pp 11328–11339. <http://proceedings.mlr.press/v119/zhang20ae.html>
- Zhong M, Liu P, Wang D, Qiu X, Huang X (2019) Searching for effective neural extractive summarization: What works and what's next. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, vol 1: Long Papers. Association for Computational Linguistics, pp 1049–1058. <https://doi.org/10.18653/v1/p19-1100>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.