EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain

Dennis Aumiller[*][†], **Ashish Chouhan**[*][†][‡] and **Michael Gertz**[†]
[†] Institute of Computer Science, Heidelberg University
[‡] School of Information, Media and Design, SRH Hochschule Heidelberg
{aumiller, chouhan, gertz}@informatik.uni-heidelberg.de

Abstract

Existing summarization datasets come with two main drawbacks: (1) They tend to focus on overly exposed domains, such as news articles or wiki-like texts, and (2) are primarily monolingual, with few multilingual datasets. In this work, we propose a novel dataset, called EUR-Lex-Sum, based on manually curated document summaries of legal acts from the European Union law platform (EUR-Lex). Documents and their respective summaries exist as cross-lingual paragraph-aligned data in several of the 24 official European languages, enabling access to various cross-lingual and lower-resourced summarization setups. We obtain up to 1,500 document/summary pairs per language, including a subset of 375 cross-lingually aligned legal acts with texts available in *all* 24 languages.

In this work, the data acquisition process is detailed and key characteristics of the resource are compared to existing summarization resources. In particular, we illustrate challenging sub-problems and open questions on the dataset that could help the facilitation of future research in the direction of domain-specific cross-lingual summarization. Limited by the extreme length and language diversity of samples, we further conduct experiments with suitable extractive monolingual and cross-lingual baselines for future work.

Code for the extraction as well as access to our data and baselines is available online at: https://github.com/achouhan93/eur-lex-sum.

1 Introduction

Despite a long history in the field of text summarization (Luhn, 1958), current systems in the area are still mainly targeted towards a few select domains. This stems in part from the homogeneity of existing summarization datasets and extraction processes: frequently, these are either collected from news articles (Over and Yen, 2004; Sandhaus, 2008; Hermann et al., 2015; Narayan et al., 2018; Grusky et al., 2018; Hasan et al., 2021) or wiki-style knowledge bases (Ladhak et al., 2020; Frefel, 2020), where alignment with supposed "summaries" is particularly straightforward. Domain outliers do exist, e.g., for scientific literature (Cachola et al., 2020) or the legal domain (Gebendorfer and Elnaggar, 2018; Kornilova and Eidelman, 2019; Manor and Li, 2019; Bhattacharya et al., 2019), but are primarily restricted to the English language or do not contain finer-grained alignments between cross-lingual documents.

Reasons for the usage of mentioned predominant domains are manifold: Data is reasonably accessible throughout the internet, can be automatically extracted, and the structure naturally lends itself to the extraction of excerpts that can be seen as a form of summarization. For news articles, short snippets (or headlines) describing the gist of main article texts are quite common. Wikipedia has an introductionary paragraph that has been framed as a "summary" of the remaining article (Frefel, 2020), whereas others utilize scholarly abstracts (or variants thereof) as extreme summaries of academic texts (Cachola et al., 2020).

For a variety of reasons, using these datasets as a training resource for summarization systems introduces (unwanted) biases. Examples include extreme lead bias (Zhu et al., 2021), focus on extremely short input/output texts (Narayan et al., 2018), or high overlap in the document contents (Nallapati et al., 2016). Models trained in such a fashion also tend to score quite well on zero-shot evaluation of datasets from similar domains, however, poorly generalize beyond immediate in-domain samples that follow a different content distribution or longer expected summary length.

Simultaneously, high-quality multilingual and cross-lingual data for training summarization systems is scarce, particularly for datasets including

[*] These authors contributed equally to this work.

more than two languages. Existing resources are often constructed in similar fashion to their monolingual counterparts (Scialom et al., 2020; Varab and Schluter, 2021) and subsequently share the same shortcomings of low data quality.
Our main contribution in this work is the construction of a novel multi- and cross-lingual corpus of reference texts and human-written summaries that extract texts from legal acts of the European Union (EU). Aside from a varying number of training samples per language, we provide a paragraph-aligned validation and test set across all 24 official languages of the European Uninon[1], which further enables cross-lingual evaluation settings.

2 Related Work

Influencing works can generally be categorized into works about EU data, or more broadly about summarization in the legal domain. Aside from that, we also compare our research to other existing multi- and cross-lingual works for text summarization.

2.1 The EU as a Data Source

Data generated by the European Union has been utilized extensively in other sub-fields of Natural Language Processing. The most prominent example is probably the Europarl corpus (Koehn, 2005), consisting of sentence-aligned translated texts generated from transcripts of the European Parliament proceedings, frequently used in Machine Translation systems due to its size and language coverage. In similar fashion to parliament transcripts, the European Union has its dedicated web platform for legal acts, case law and treaties, called EUR-Lex (Bernet and Berteloot, 2006)[2], which we will refer to as the *EUR-Lex platform*. Data from the EUR-Lex platform has previously been utilized as a resource for extreme multi-label classification (Loza Mencía and Fürnkranz, 2010), most recently including an updated version by Chalkidis et al. (2019a,b). In particular, the MultiEURLEX dataset (Chalkidis et al., 2021) extends the monolingual resource to a multilingual one, however, does not move beyond the classification of EuroVoc labels. To our knowledge, document summaries of legal acts from the platform have recently been used as a monolingual English training resource for summarization systems (Klaus et al., 2022).

2.2 Processing of Long Legal Texts

Recently, using sparse attention, transformer-based models have been proposed to handle longer documents (Beltagy et al., 2020; Zaheer et al., 2020a). However, the content structure is not explicitly considered in current models. Yang et al. (2020) proposed a hierarchical Transformer model, SMITH, that incrementally encodes increasingly larger text blocks. Given the lengthy nature of legal texts, (Aumiller et al., 2021) investigate methods to separate content into topically coherent segments, which can benefit the processing of unstructured and heterogeneous documents in long-form processing settings with limited context. From a data perspective, Kornilova and Eidelman (2019) propose BillSum, a resource based on US and California bill texts, spanning between approximately 5,000 to 20,000 characters in length. For the aforementioned English summarization corpus based on the EUR-Lex platform, Klaus et al. (2022) utilize an automatically aligned text corpus for fine-tuning BERT-like Transformer models on an extractive summarization objective. Their best-performing approach is a hybrid solution that prefaces the Transformer system with a TextRank-based pre-filtering step.

2.3 Datasets for Multi- or Cross-lingual Summarization

For Cross-lingual Summarization (XLS), Wang et al. (2022b) provide an extensive survey on the currently available methods, datasets, and prospects. Resources for XLS can be divided into two primary categories: synthetic datasets and web-native multilingual resources. For the former, samples are created by directly translating summaries from a given source language to a separate target language. Examples include English-Chinese (and vice versa) by Zhu et al. (2019), and an English-German resource (Bai et al., 2021). Both works utilize news articles for data and neural MT systems for the translation. In contrast, there also exist web-native multilingual datasets, where both references and summaries were obtained primarily from parallel website data. Global Voices (Nguyen and Daumé III, 2019), XWikis (Perez-Beltrachini and Lapata, 2021), Spektrum (Fatima and Strube, 2021), and CLIDSUM (Wang et al., 2022a) represent instances of datasets for the news, encyclopedic, and dialogue domain, with differing numbers

[1] `https://eur-lex.europa.eu/content/help/eurlex-content/linguistic-coverage.html`, last accessed: 2022-06-15

[2] most recent URL: `https://eur-lex.europa.eu`, last accessed: 2022-06-15

of supported languages.

We have previously mentioned some of the multilingual summarization resource where multiple languages are covered. MLSUM (Scialom et al., 2020) is based on news articles in six languages, however, without cross-lingual alignments. Similarly without alignments, but larger in scale, is MassiveSum (Varab and Schluter, 2021). XL-Sum Hasan et al. (2021) does provide document-aligned news article, in 44 distinct languages, extracted data from translated articles published by the BBC. In particular, their work also provides translations in several lower-resourced Asian languages. WikiLingua (Ladhak et al., 2020) borders the multi- and cross-lingual domain; some weak alignments exist, but only for English references, and not between languages themselves.

3 The EUR-Lex-Sum Dataset

We present a novel dataset based on available multilingual document summaries from the EUR-Lex platform. The final dataset, which we title "*EUR-Lex-Sum*", consists of up to 1,500 document/summary pairs per language. For comparable validation and test splits, we identified a subset of 375 cross-lingually aligned legal acts that are available in all 24 languages. In this section, the data acquisition process is detailed, followed by a brief exploratory analysis of the documents and their content. Finally, key intrinsic characteristics of the resource are compared with relation to existing summarization resources. In short, we find that the combination of human-written summaries coupled with comparatively long source *and* summary texts makes this dataset a suitable resource for evaluating a less common summarization setting, especially for long-form tasks.

3.1 Dataset Creation

The EUR-Lex platform provides access to various legal documents published by organs within the European Union. In particular, we focus on currently enforced EU legislation (legal acts) for the 20 domains from the EUR-Lex platform.[3] From the mentioned link, direct access to lists of published legal acts associated with a particular domain is available, which forms the starting point for our later crawling step. Notably, each of these domains also provides a diverse set of specific keywords, topics and regulations, which even within the dataset provide a high level of diversity.

A legal act is uniquely identified by the so-called Celex ID, composed of codes for the respective sector, year and document type. The ID is consistent across all 24 languages, which makes it possible to align articles on a document level. Across all 20 domains, the website reports a total of 26,468 legal acts spanning from 1952 until 2022. However, as there is a probability of a particular legal act being assigned to multiple domains, approximately 22,000 unique legal acts can be extracted from the platform. We do not consider EU case law and treaties, which are also available through the EUR-Lex platform, but in other document formats.

3.1.1 Crawling

The web page of a particular legal act contains the following page content relevant for a summarization setting: 1. The published text of the particular legal act in various file formats, 2. metadata information about the legal acts, such as published year, associated treaties, etc., 3. links to the content pages in other official languages, and 4. if available, a link to an associated summary document.

This work contributes to preparing a dataset with the legal act content and their respective summaries in different languages. Therefore, crawling over the entirety of published legal acts gives access to all relevant information needed to extract source and summary text pairs. Since a single legal act requires 50 individual web requests to extract files across all languages, we have a total of around 5.5 million access requests, distributed across the span of a month between May and June 2022. We dump the content of all accessed acts in a local Elasticsearch instance, and separately mark documents without existing associated summaries. This allows the resource to be continually updated in the future without re-crawling documents that do not have available summaries.

3.1.2 Filtering

For further processing, we filter the documents available through our offline storage. First, some article texts may only be available as scanned (PDF) documents, which compromises text quality and is therefore discarded. For the most consistent representation, we choose to limit ourselves to articles present in an HTML document, with further advantages explained in Section 4.1. Availability of

[3] https://eur-lex.europa.eu/browse/directories/legislation.html, last accessed: 2022-06-21

HTML documents generally correlates with the publishing year, see Section 4.2, presumably due to the emergence of the world wide web during the 1990s. Similarly, a document is not required to have an associated summary, limiting sample pairs' availability. A full distribution of available HTML sample pairs can be found in Figure 4. We could not identify any particular reasoning behind what documents do have summaries and which do not.

More problematic is the fact that between 20-30% of the available summaries (depending on the language) are associated with *several* source documents, essentially turning this into a multi-document summarization setting. Since this work focuses exclusively on single document summarization, we pair the summary with the longest associated reference document to maximize availability. Table 1 details the impact of considering only the longest document in terms of n-gram novelty; we observe a consistent increase of novel n-grams by about 5 percentage points over the subset of single-reference documents. While the concatenation of all relevant reference documents would eliminate any difference in n-gram overlap between the summary and reference texts, having a single reference document conserves the correct processing of lead biases over alternatives that aggregate several texts. Further, concatenation leads to ambiguous text orderings, which may change summarization outcomes based on different aggregation strategies. However, the subset of these multi-document samples could be a challenging problem based on our available corpus that may be explored in a separate context for future work.

Finally, we filter out all document pairs where the reference text is shorter than the input document. This occurs only for multi-document summary pairs, where sometimes several short acts are aggregated into a single summary.

After filtering out invalid samples, between 391 (Irish) to 1,505 (French) documents remain; the full list of samples broken down by language can be found in the Appendix in Table 5. Across all languages, we manage to extract 31,987 pairs.

3.1.3 Data Split

To ensure a suitable (and comparable) validation and test split across different languages, all documents having sample pairs available in 24 languages (375 total) are taken out of the available respective subsets. Of the 375 documents, 187 samples are randomly selected into the validation set, and the remaining 188 are taken as the corresponding test set. All other documents are assigned to the language-dependent training sets. No guarantee for cross-lingual availability is provided for the training set, however, most documents do appear in several of the languages. In particular, We will use these filtered data splits for future experiments in this paper unless explicitly mentioned otherwise.

	n-gram novelty			
Subset	1-gram	2-gram	3-gram	4-gram
All samples	42.25	64.07	77.34	83.73
Single-ref subset	41.74	63.52	76.87	83.33
Longest	46.77	68.83	81.44	87.18
Concatenated	41.03	63.06	76.38	82.77

Table 1: Comparison of n-gram novelties for the English subset, differentiating by the number of reference documents. *Longest* considers the subset of multi-reference documents with only the longest document as a reference; *Concatenated* uses the concatenation of all associated references.

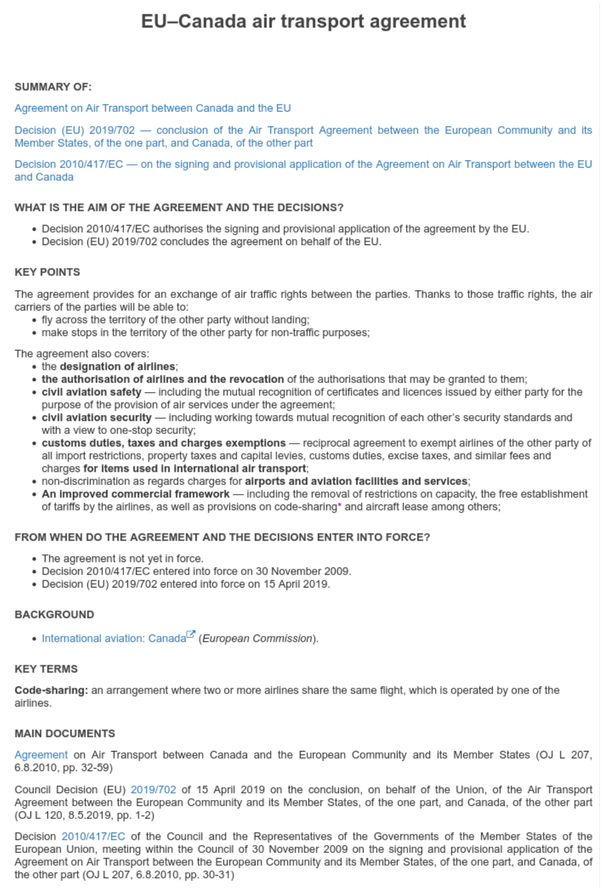
4 Exploratory Analysis

An exploratory analysis of the dataset is conducted to confirm the resource's viability for automatic summarization and overall data quality. Aside from a qualitative view of the resource and an analysis of the temporal distribution of our samples, we provide a comprehensive look at intrinsic metrics commonly used for summarization datasets.

4.1 Data Quality

Documents of the EU are generally held to a high standard, and the legal acts are no exception. This also extends to the summaries, which follow a particular set of guidelines for their creation process.[4] In particular, guidelines for drafting summary texts are detailed in Technical Annex I, which specify several key instructions for generating human-written summaries of an underlying legal act. Most prominently, they recommend a target length for key point summaries between 500-700 words and formulate a template structure for the overall text outline. An example of a typical summary structure can be seen in Figure 1. Aside from the key points, this includes, e.g., references to the main documents or specific act-related key phrases. We want to highlight that the generation guidelines changed

[4] https://etendering.ted.europa.eu/cft/cft-documents.html?cftId=6490, last accessed: 2022-06-15.



EU–Canada air transport agreement

SUMMARY OF:

Agreement on Air Transport between Canada and the EU

Decision (EU) 2019/702 — conclusion of the Air Transport Agreement between the European Community and its Member States, of the one part, and Canada, of the other part

Decision 2010/417/EC — on the signing and provisional application of the Agreement on Air Transport between the EU and Canada

WHAT IS THE AIM OF THE AGREEMENT AND THE DECISIONS?

- Decision 2010/417/EC authorises the signing and provisional application of the agreement by the EU.
- Decision (EU) 2019/702 concludes the agreement on behalf of the EU.

KEY POINTS

The agreement provides for an exchange of air traffic rights between the parties. Thanks to those traffic rights, the air carriers of the parties will be able to:

- fly across the territory of the other party without landing;
- make stops in the territory of the other party for non-traffic purposes;

The agreement also covers:

- the **designation of airlines**;
- **the authorisation of airlines and the revocation** of the authorisations that may be granted to them;
- **civil aviation safety** — including the mutual recognition of certificates and licences issued by either party for the purpose of the provision of air services under the agreement;
- **civil aviation security** — including working towards mutual recognition of each other's security standards and with a view to one-stop security;
- **customs duties, taxes and charges exemptions** — reciprocal agreement to exempt airlines of the other party of all import restrictions, property taxes and capital levies, customs duties, excise taxes, and similar fees and charges **for items used in international air transport**;
- non-discrimination as regards charges for **airports and aviation facilities and services**;
- **An improved commercial framework** — including the removal of restrictions on capacity, the free establishment of tariffs by the airlines, as well as provisions on code-sharing* and aircraft lease among others;

FROM WHEN DO THE AGREEMENT AND THE DECISIONS ENTER INTO FORCE?

- The agreement is not yet in force.
- Decision 2010/417/EC entered into force on 30 November 2009.
- Decision (EU) 2019/702 entered into force on 15 April 2019.

BACKGROUND

- International aviation: Canada (*European Commission*).

KEY TERMS

Code-sharing: an arrangement where two or more airlines share the same flight, which is operated by one of the airlines.

MAIN DOCUMENTS

Agreement on Air Transport between Canada and the European Community and its Member States (OJ L 207, 6.8.2010, pp. 32-59)

Council Decision (EU) 2019/702 of 15 April 2019 on the conclusion, on behalf of the Union, of the Air Transport Agreement between the European Community and its Member States, of the one part, and Canada, of the other part (OJ L 120, 8.5.2019, pp. 1-2)

Decision 2010/417/EC of the Council and the Representatives of the Governments of the Member States of the European Union, meeting within the Council of 30 November 2009 on the signing and provisional application of the Agreement on Air Transport between the European Community and its Member States, of the one part, and Canada, of the other part (OJ L 207, 6.8.2010, pp. 30-31)

Figure 1: Summary of legal act with Celex ID 32019D0702. Visible are several distinct sections, with the majority of the document describing key points of the underlying legal acts. This particular summary aggregates content from several legal acts, of which we consider the longest one as the reference document.

over time. Since we do not have access to previous versions of the guidelines, we manually probed comparisons between older and newer documents, which exposed a highly structural similarity despite changes in guidelines.

The published documents and summaries offer further peculiarities in both their content structure as well as the creation process: First, the multilingual versions of both documents and summaries are always translated from the original English legal act (or English summary thereof),[5] which ensures strict content similarity of the same text across all available languages. Second, due to their HTML representation, it is possible to extract *paragraph-aligned* texts between language-specific versions. This is a well-known property of EU-level data, most notably exploited in the Europarl corpus (Koehn, 2005) for automatic alignments of machine translation training data. We similarly maintain this structure during the extraction process to use it at later stages, e.g., for more informed evaluation setups or cross-lingual pre-training.

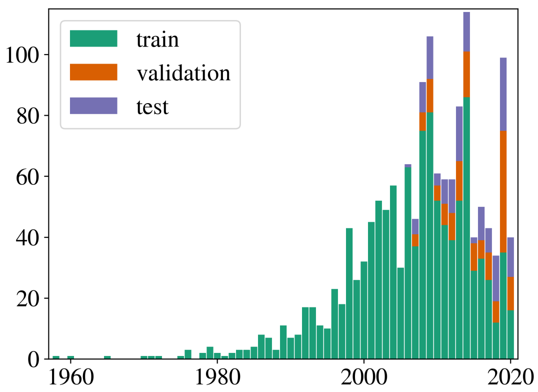


Figure 2: Distribution of the publishing year of unique legal acts included in the final dataset. Availability of documents increases after 1990.

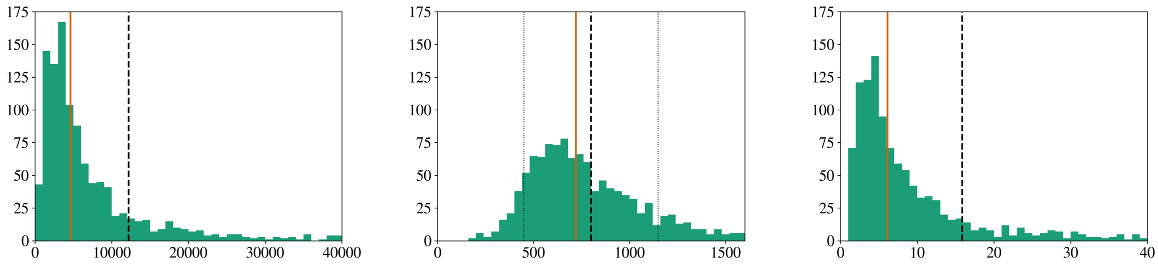
4.2 Temporal Distribution

Figure 2 displays the distribution of filtered documents by the year of publication. The amount of available samples increases after 1990, which likely coincides with more member states joining, as well as a shift to digital archiving (compared to OCR scans of PDF documents, which are excluded from our corpus). Compared to other European resources, such as Multi-EURLex (Chalkidis et al., 2021), a lesser topical shift is expected in our resource, simply due to a more limited time frame. Notably, we also include the distribution by dataset split and observe an even stronger bias towards more recent legal acts for validation and test sets. This is a natural consequence of the requirement for validation and test sets that legal acts be present in all 24 languages, which includes more recently added official language, such as Croatian (added in 2013) or Irish (added in 2022). We also want to mention that amendments to both reference and summary texts might be added (or revised) several years after their original publication, which is not reflected in our analysis.

4.3 Document Structure

An example of the content structure of a document summary is provided in Figure 1. The formatted text reveals significant sections of the summary,

[5] This has been confirmed by the Publications Office of the European Union in private correspondence.



(a) Reference tokens

(b) Summary tokens

(c) Token compression ratio

Figure 3: Histogram of the English training set, comparing article token lengths. Displayed are the distribution for references (left), summaries (center), and compression ratios (right). Vertical lines show median length (continuous orange), mean length (dashed black), and standard deviation (dotted black lines). The latter exceeds display limits for reference length and compression ratio. Ranges are limited to the 95th length percentile for legibility.

where the majority is taken up by free text describing the key goals and highlights of the sub-points within a longer legal act. It further describes which legal act (or several acts) are associated with the summary. As previously described, we limit ourselves to the longest associated legal act for a summary associated with several acts.

While we provide raw text for the extracted legal act document in the proposed resource, example document in Figure 1 reveals a potential use case of semi-structured visual information from HTML tags (e.g., headline descriptors or bullet lists), which could be used for a fine-grained distinction between different content parts. In our preliminary experiments, we found that the used HTML tags for content elements can vary significantly between different legal acts (e.g., using modified `div` containers instead of `H3` for subheadings) and therefore keep the inclusion of such features for future work.

4.4 Summarization-related Dataset Metrics

We adopt metrics from prior work to automatically analyze summarization datasets (Grusky et al., 2018; Zhong et al., 2019; Bommasani and Cardie, 2020). Our corpus reveals a high degree of abstractivity, which is surprising given the enormous length of input texts.

4.4.1 Length Distribution

Based on the fact mentioned in Section 4.1 that documents are created as translations from the English original, we focus more on the distribution of legal acts and their summary lengths in English as a representative language. A more exhaustive overview can be found in the Appendix in Table 5, which gives more insight into language-specific length variations due to document availability, or simply morphological/syntactic differences, e.g., compound words.

Histogram plots in Figure 3 show a Zipfian distribution for reference text lengths, with a mean of around 12,000 tokens; however, we also observe an exceptionally large standard deviation due to extreme outliers, mentioned in Table 2. In contrast, summary lengths exhibit closer to a normal distribution, which matches the guideline document's suggested length of 500-700 words. The observed mean is slightly higher at around 800 tokens, which can be attributed to document overhead not counting towards the actual summarizing content, such as referenced documents and separately highlighted key concepts. However, we also observe extreme outliers for summary texts (cf. Table 2).

4.4.2 Compression Ratio

We follow the definition of an unrestricted compression ratio (Grusky et al., 2018), dividing the (token) length of an article by the associated summary (token) length. This carries the same semantic value as inverse definitions of compression ratio, such as used by Bommasani and Cardie (2020). When looking at the token-level compression ratio displayed in Figure 3, a comparatively high mean is observed, despite extremely long summary documents. Comparing compression ratios reported by (Zhong et al., 2019) for news-based datasets indicates that EUR-Lex-Sum has a mean compression ratios similar to the CNN/DailyMail (Hermann et al., 2015) and NYT (Sandhaus, 2008) corpora.

4.4.3 n-gram Novelty

To provide insight into the abstractiveness of gold summaries, we follow Narayan et al. (2018) in an-

	Reference tokens		**Summary tokens**		**Comp.**	**% novel n-grams in summary**			
	Min	**Max**	**Min**	**Max**	**Ratio**	**1-gram**	**2-gram**	**3-gram**	**4-gram**
Train	385	1,087,217	173	3021	16 ± 62	44.10	65.97	78.85	84.96
Val.	1,143	199,405	354	5136	18 ± 17	36.65	58.23	72.74	79.96
Test	1,544	403,319	369	2987	18 ± 20	36.78	58.46	72.83	80.07

Table 2: Complementary dataset properties (min and max token lengths of both reference texts and summaries), as well as compression ratio, across subsets. We further report novelty n-gram shares in the gold summary. Values are computed on the English subset splits.

alyzing the fraction of n-grams not present in the original reference article. This metric is similar to content coverage metrics used by Grusky et al. (2018) or Zhong et al. (2019). When comparing novelty n-grams reported in Table 2, it should be noted that this slightly overestimates the real value. This can be attributed to our use of whitespace tokenization, which may cause more n-grams due to decreased tokenization accuracy; we further discuss the tokenization choice in Section 5.

5 Experiments

As a reference for future work building on top of this dataset, we provide a set of suitable baselines and discuss limitations of methods and data. Notably, there are considerable challenges in constructing baseline runs with popular algorithms on this dataset if trying to cover all languages.

Primarily, even summary lengths exceed input limitations of popular abstractive neural models based on transformer architectures; these systems are generally limited to 512 (subword) tokens (Lewis et al., 2020; Xue et al., 2021), and even length-focused alternatives generally boast only up to 4096 tokens (Beltagy et al., 2020; Zaheer et al., 2020b), which is well below the median length of reference texts and prevents us from training systems without further (manual) alignments provided on chunks of the input text.

Less obvious, but no less problematic is the availability of tokenizers or sentence splitting methods in popular NLP libraries, affecting several languages in our corpus (for a more in-depth list of supported languages by library, see Appendix Table 5). This inherently prevents fair sentence-level evaluation (or extraction), as system performance is not guaranteed for underrepresented languages.

Aside from a set of extractive baselines, we further evaluate a cross-lingual scenario in which summaries for the English reference text are generated and then translated into the target languages. The hypothesis is that this provides insight into limitations of existing XLS systems discussed in Section 2.3 and also represents more realistic deployment scenarios where XLS systems can be utilized as supportive summarizers for monolingual input texts.

5.1 Zero-shot Extractive Baselines

One popular traditional algorithm for generating extractive summaries is LexRank (Erkan and Radev, 2004). We utilize a modified variant of LexRank that uses multilingual embeddings generated by sentence-transformers (Reimers and Gurevych, 2019, 2020) to compute centrality. Given the previously mentioned limitations of sentencizing input texts, we chunk the text based on existing paragraph separators (refer to Figure 1), and treat those segments as inputs to our baseline setup. Notably, this method does not require any form of fine-tuning or language adoption and works as a zero-shot domain transferred extractive model, which makes it preferable over methods such as SummaRuNNer (Nallapati et al., 2017) or extractive BERT summarizers,[6] which require training on (automatically extracted) alignments.

To determine the output summary length, we calculate the average paragraph-level compression ratio on the language's training set, and then multiply this value with the reference document's number of paragraphs to obtain a target length.

For evaluation, we rely on ROUGE scores (Lin, 2004) with disabled stemming to conserve comparability between languages. We acknowledge that this is not a comprehensive measure and has distinctive shortcomings, but works fairly well at the paragraph level, as such units generally preserve both factual consistency and fluency.

Due to the paragraph-level consistency of generated summaries, this is a fairly strong baseline. Importantly, ROUGE scores remain consistent for lan-

[6] e.g., `bert-extractive-summarizer`

	Validation			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
English	25.99	13.34	13.30	26.68	13.65	13.58
French	32.18	18.03	15.15	32.35	18.00	15.16
German	26.00	13.12	12.24	26.72	13.75	12.56
Spanish	27.04	16.43	14.75	28.34	17.12	15.23
Italian	27.29	14.01	12.63	28.57	14.24	12.90
Portuguese	30.12	17.17	15.08	30.67	17.20	15.20
Dutch	29.07	14.92	14.66	29.62	14.76	14.73
Danish	28.78	13.90	13.14	29.22	13.86	13.19
Greek	24.42	9.77	15.46	24.79	9.45	15.46
Finnish	26.40	11.88	11.87	26.49	11.68	11.80
Swedish	30.25	15.40	14.27	30.67	15.47	14.35
Romanian	35.69	16.08	14.90	34.75	15.16	14.59
Hungarian	33.71	19.53	15.49	34.55	19.69	15.64
Czech	30.96	16.65	14.16	31.86	16.76	14.32
Polish	28.47	14.42	12.68	28.88	14.42	12.73
Bulgarian	26.36	9.15	16.54	25.58	8.40	16.13
Latvian	31.24	15.55	12.99	31.73	15.77	13.15
Slovene	26.75	12.25	11.64	27.19	12.34	11.79
Estonian	26.33	11.64	11.84	26.39	11.41	11.66
Lithuanian	26.79	12.43	11.44	26.76	12.45	11.59
Slovak	30.30	15.04	13.14	30.65	14.94	13.14
Maltese	29.71	14.55	12.73	30.51	14.62	12.86
Croatian	33.50	13.46	13.50	32.64	12.76	13.29
Irish	43.66	18.72	15.86	41.93	17.16	15.25

Table 3: Extractive summarization baseline with modified LexRank. We report ROUGE F1 scores for both the validation and test splits.

guages between the validation and test set, although we do observe some languages with outlier performance: For Greek text, the model likely struggles with the representation of non-arabic subwords, but still performs decently well at the ROUGE-L level. Otherwise, Irish has unexpectedly high ROUGE scores, which we were unable to explain. This is especially surprising given the fact that the language is not even one officially supported by the multilingual embedding model used for this experiment.

5.2 Cross-lingual Baselines

As a baseline for XLS, we provide a simple two-step translate-then-summarize pipeline (Wang et al., 2022b). To generate summaries on longer contexts, we utilize a model based on the Longformer Encoder Decoder (LED) architecture (Beltagy et al., 2020), precisely a checkpoint previously fine-tuned on the English BillSum corpus (Kornilova and Eidelman, 2019). Translation from English to target languages is done with OPUS-MT (Tiedemann and Thottingal, 2020). To deal with long documents exceeding the particular model's window size, we greedily chunk text if necessary. To represent an upper limit of performance, we compare a translate-then-summarize setup from English to Spanish, which can be regarded as one of the language pairs with the highest MT performance, due to data availability and linguistic similarity of the source and target language.

	Validation			Test		
	R-1	R-2	R-L	R-1	R-2	R-L
LED	31.67	13.00	16.17	31.14	13.01	16.20
LexRank-EN	39.42	20.03	18.53	39.44	20.02	18.73
LexRank-ES	27.04	16.43	14.75	28.34	17.12	15.23
Oracle	52.84	39.79	43.87	54.55	41.01	45.06

Table 4: Cross-lingual summarization setup for English-Spanish. We report ROUGE F1 scores for both the validation and test splits on the Spanish subset.

As baselines, we provide translations of the English gold summaries into the target language (again with the Opus MT model), as well as a translation of the extractive LexRank summary from the previous experiment. Results seen in Table 4 are surprising: While the abstractive model seems to improve over the purely Spanish-based LexRank summary (LexRank-ES) by a significant margin, it turns out that translating the English LexRank baseline drastically *improves* results in terms of ROUGE scores. We assume that this is related to truncation and re-phrasing happening during the translation step.

5.3 Open Problems

The most obvious problem for this dataset is the extreme length, and also length disparity between documents. This is especially apparent when comparing the length to average samples in CNN/DailyMail (Hermann et al., 2015), where the mean article length is about 16 times shorter; this makes content selection significantly more challenging.

Secondly, incorporating hierarchical information about the reference text could greatly improve context relevance in such extensive settings. However, this is not only restricted to the reference document, but could also be considered for the (hierarchical) construction of long-form summary texts. Given that previous datasets do not come with such long output samples, this has to our knowledge not been previously tackled in the literature.

Ultimately, the question of equal coverage for lesser-resourced languages is also not fully answered. While we attempt to treat languages in our dataset equally, this comes with its particular set of challenges and performance hits in highly available languages.

6 Conclusion and Future Work

Throughout this work, we have detailed the creation of a new multilingual corpus for text summarization, based on legal acts from the European Union. We further provided a more detailed analysis of the underlying data and sample quality and hypothesized potential applications to open problems in the communtiy, such as long-form summarization or cross-lingual application scenarios. Our dataset is publicly available on the web, and comes with a set of monolingual extractive baselines that provide suitable reference points for any future work in this direction.

In particular, we intend to focus on exploiting the structure of summaries for a more guided generation of output texts. Especially for extremely long legal texts, template structures could be utilized. On a more general level, we expect that progress in long-form models is required to achieve remotely sensible results on extreme-length generative tasks. Alternative approaches in the meantime could include aspect-driven methods for building summaries in an iterative fashion.

Finally, on top of the static snapshot presented in this work, we are also working towards a continually updated data repository of this resource, which would then include newly added texts (or summaries) for EU texts.

7 Limitations

While our work considers comparatively high-quality data samples, there still remain some assumptions about the underlying text sources, which lead to some of the following limitations:

1. Documents themselves (both sources and summaries) may link to external articles or related regulations for further information. Some of the linked documents might indeed contain relevant contextual information, but are as such not considered in our version.

2. On a similar note, we mentioned that some summaries aggregate content from *several legal acts*, as outlined in Section 3.1.2; this is not considered in full at the current stage and might cause limitations.

3. Legal acts may exist in several iterations, drafted up at different points in time. To the best of our knowledge, we extracted the most recent version and its associated summary.

4. For evaluation of generated summarization quality, we provide n-gram-based ROUGE scores, which have previously been argued to poorly reflect particular aspects, e.g., factual consistency or fluency. Given this, baseline performance should be taken in clear context for future work.

Broader Impact & Ethical Issues

With the release of our data as a public resource, we want to touch on the potential ethical implications of this release: As our data is already available (though much more inaccessible) through the EUR-Lex platform, we do not see any ethical concerns in a repurposed and bundled release of this dataset from such a standpoint. To our knowledge, human-written reference and summary texts, as well as the accompanying translations into the European languages, have undergone review within instances of the European Union, leading to no clear concerns in data quality, especially with respect to potential privacy violations or harmful text content.

However, we acknowledge that this resource reinforces a certain availability bias towards European languages, which needs to be acknowledged by follow-up work. On the other hand, we believe that the release of resources including underrepresented languages, for example, Irish and Maltese, outweighs this concern and fosters future research in a more multilingual way.

Acknowledgements

We want to especially thank Reviewer 1 for their constructive feedback and suggestion of further experiments in their review.

References

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, pages 2–11. ACM.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hélène Bernet and Pascale Berteloot. 2006. Eur-lex: A multilingual on-line website for european union law. *International Review of Law Computers & Technology*, 20(3):337–339.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, volume 11437 of *Lecture Notes in Computer Science*, pages 413–428. Springer.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019a. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019b. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Mehwish Fatima and Michael Strube. 2021. A novel Wikipedia based dataset for monolingual and cross-lingual summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 39–50, Online and in Dominican Republic. Association for Computational Linguistics.

Dominik Frefel. 2020. Summarization corpora of Wikipedia articles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6651–6655, Marseille, France. European Language Resources Association.

Christoph Gebendorfer and Ahmed Elnaggar. 2018. Legal jrc-acquis sum – text summarization corpus. In *Technical University of Munich*. (Date accessed: 21.06.2022).

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Svea Klaus, Ria Van Hecke, Kaweh Djafari Naini, Ismail Sengor Altingovde, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. 2022. Summarizing legal regulatory documents using transformers. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2426–2430. ACM.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Eneldo Loza Mencía and Johannes Fürnkranz. 2010. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts – Where the Language of Law Meets the Law of Language*, 1 edition, volume 6036 of *Lecture Notes in Artificial Intelligence*, pages 192–215. Springer-Verlag.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Paul Over and James Yen. 2004. An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus, LDC2008T19.

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

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. A survey on cross-lingual summarization. *arXiv preprint arXiv:2203.12515*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 1725–1734, New York, NY, USA. Association for Computing Machinery.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020a. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020b. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89, Hong Kong, China. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1462–1471. ACM.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

A Language-specific Distributions

Figure 4 displays available articles before filtering. Notably, the documents need not be subsets of one another, meaning the French document IDs might differ from English ones. Table 5 further compares the availability of language-specific articles before and after filtering, to provide an insight into the number of removed documents. The same table also provides a more concise overview of supported languages in popular frameworks, as well as an extension of statistics reported in Table 2 for the language-specific training sets. To illustrate cross-lingual presence of Celex IDs, we plot the inverse availability distribution (sample is available in *at least k languages*) in Figure 5. Around 84% of the samples are available in 20 or more languages.

B Implementation Details for Baselines

For extractive monolingual models, we use the checkpoint "paraphrase-multilingual-mpnet-base-v2"[7] without any further fine-tuning, using version

[7] model configuration: `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2/blob/main/config.json`, last accessed: 2022-06-23

	No. articles		**Availability**			**Article token length**		**Comp.**	**n-gram novelty**	
Language	**before**	**after**	**S-T**	**spacy**	**nltk**	**Reference**	**Summary**	**ratio**	**1-gram**	**2-gram**
English (en)	1,974	1,504	✓	✓	✓	12206 ± 42429	799 ± 349	16 ± 62	44.10	65.97
French (fr)	1,969	1,505	✓	✓	✓	13192 ± 43950	892 ± 395	16 ± 63	45.07	64.13
German (de)	1,966	1,490	✓	✓	✓	11144 ± 41061	748 ± 330	16 ± 68	44.85	66.95
Spanish (es)	1,964	1,487	✓	✓	✓	13581 ± 44574	932 ± 420	15 ± 57	44.76	61.51
Italian (it)	1,867	1,403	✓	✓	✓	13152 ± 44641	845 ± 370	16 ± 67	44.77	67.00
Portuguese (pt)	1,845	1,376	✓	✓	✓	12629 ± 29921	896 ± 391	14 ± 38	43.84	64.00
Dutch (nl)	1,844	1,376	✓	✓	✓	13233 ± 44638	834 ± 362	17 ± 69	44.41	65.86
Danish (da)	1,843	1,377	✓	✓	✓	11947 ± 43155	717 ± 308	18 ± 71	46.96	68.27
Greek (el)	1,837	1,366	✓	✓	✓	13609 ± 45411	863 ± 369	17 ± 64	44.86	66.70
Finnish (fi)	1,825	1,366	✓	✓	✓	9792 ± 41021	575 ± 247	18 ± 93	53.41	77.26
Swedish (sv)	1,822	1,362	✓	✓	✓	10796 ± 26923	718 ± 305	15 ± 40	46.74	69.62
Romanian (ro)	1,817	1,353	✓	✓	✗	13646 ± 45644	826 ± 356	17 ± 67	45.42	67.80
Hungarian (hu)	1,813	1,336	✓	✗	✗	12230 ± 46764	702 ± 298	19 ± 84	53.23	75.68
Czech (cs)	1,812	1,359	✓	?	✓	12469 ± 46640	715 ± 307	18 ± 77	46.75	71.89
Polish (pl)	1,811	1,353	✓	✓	✓	11560 ± 33296	739 ± 324	16 ± 48	46.69	71.01
Bulgarian (bg)	1,792	1,332	✓	✗	✗	13397 ± 45578	819 ± 350	17 ± 69	47.00	68.44
Latvian (lv)	1,790	1,334	✓	?	✗	11841 ± 46552	670 ± 289	19 ± 83	50.23	74.55
Slovene (sl)	1,789	1,332	✓	✗	✓	11357 ± 32842	712 ± 305	16 ± 48	47.28	71.57
Estonian (et)	1,788	1,332	✓	✗	✓	10778 ± 45157	581 ± 249	20 ± 94	52.20	77.46
Lithuanian (lt)	1,788	1,335	✓	?	✓	11943 ± 46673	669 ± 290	19 ± 88	47.79	74.00
Slovak (sk)	1,788	1,325	✓	?	✗	11600 ± 32968	729 ± 319	16 ± 47	48.20	73.42
Maltese (mt)	1,770	1,315	✗	✗	✗	12711 ± 48156	685 ± 299	20 ± 85	54.77	81.43
Croatian (hr)	1,762	1,278	✓	?	✗	10051 ± 19390	712 ± 307	14 ± 28	48.62	72.61
Irish (ga)	427	391	✗	?	✗	28152 ± 63360	948 ± 385	46 ± 137	45.89	70.38

Table 5: Supplementary statistics of the EUR-Lex-Sum corpus across languages. We list the total number of available articles (before and after filtering), and whether a particular language is supported by `sentence-transformers` multilingual models ("*S-T*"), or has available language-specific models in `spaCy` (Honnibal et al., 2020) or `nltk` (Bird et al., 2009), respectively. "?" indicates potential support through general-purpose models with uncertain segmentation quality. We also provide abriged statistics along the lines of Figure 3 and Table 2 for the training partition of all languages.

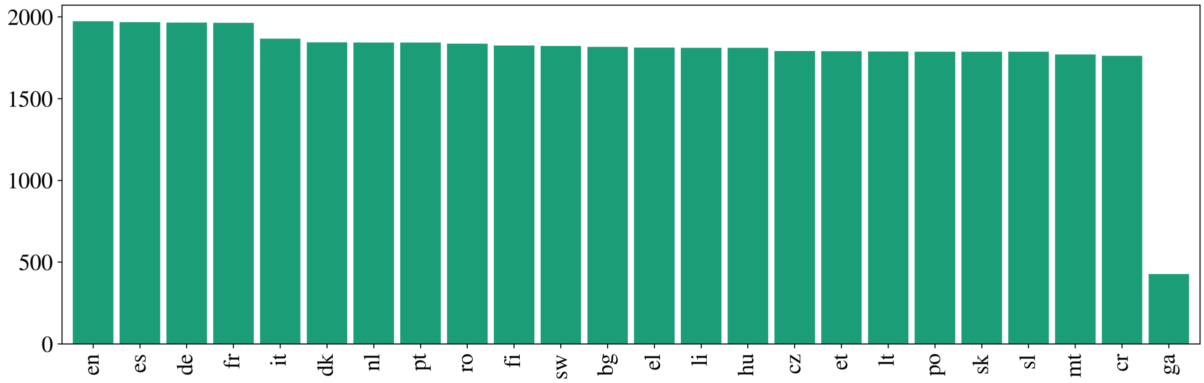


Figure 4: The number of all crawled document/summary pairs across the 24 official EU languages *before* filtering. Irish exhibits a greatly limited availability due to its recent addition as an official language.

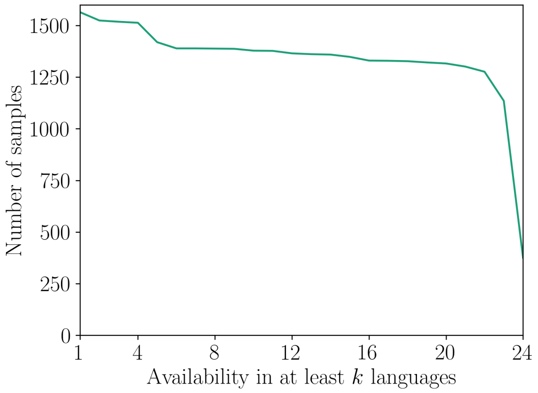


Figure 5: Celex IDs present in at least k languages.

2.1.0 of `sentence-transformers`. We do use a slightly modified version of their LexRank implementation to avoid a bug preventing the power method from converging due to "negative likelihoods". This can be fixed by normalizing similarity scores to strictly positive values.

The abstractive LEDBill model[8] was used through the `pipeline` feature available in Huggingface Transformers (Wolf et al., 2020), version 4.18. We use greedy decoding for text generation and chunk text into blocks of approximately 4096 tokens, where we then concatenate the output summaries of consecutive sections. A similar setup was used for translation, where we use the Opus MT models for respective language pairs (`HelsinkiNLP/opus-mt-<src>-<tgt>`) (Tiedemann and Thottingal, 2020), although the context size used for chunking is 500 subword tokens (to account for model-specific padding). We refer to the model card for configuration hyperparameters of LEDBill[9] and Opus MT[10]. Parameter counts for all three neural models can be found in Table 6.

For the computation of ROUGE scores, we utilize the implementation by Google Research[11], with stemming disabled.

For GPU inference, we use a machine with a single Nvidia Titan RTX with 24 GB GPU VRAM and 64 GB RAM. Obtaining results for the LexRank baselines on both the test and validation set takes less than 2.5 minutes on average for the validation and test samples (375 total generated summaries). In comparison, the generation with LEDBill takes approximately 12 hours per 375 validation/test samples. Computationally speaking, translations lie somewhere in between the previous settings, taking around 20 minutes to compute.

Model	**Parameters**
S-T	278MM
LEDBill	162MM
Opus MT	78MM

Table 6: Approximate parameter count for utilized neural systems. "S-T" represents the sentence-transformers model used for computing sentence embbeddings in the modified LexRank baseline.

[8] `https://huggingface.co/d0r1h/LEDBill`, last accessed: 2022-06-23

[9] `https://huggingface.co/d0r1h/LEDBill/blob/main/config.json`, last accessed: 2022-06-23

[10] `https://huggingface.co/Helsinki-NLP/opus-mt-en-es/blob/main/config.json`, last accessed: 2022-06-23

[11] `https://pypi.org/project/rouge-score/`; version 0.0.4, last accessed: 2022-06-23