# EUROPA: A Legal Multilingual Keyphrase Generation Dataset

**Olivier Salaün, Frédéric Piedboeuf, Guillaume Le Berre**
**David Alfonso Hermelo, Philippe Langlais**
RALI, DIRO, Université de Montréal, Canada
{olivier.salaun, frederic.piedboeuf,
guillaume.le.berre, philippe.langlais}@umontreal.ca
david.alfonso.hermelo@gmail.com

## Abstract

Keyphrase generation has primarily been explored within the context of academic research articles, with a particular focus on scientific domains and the English language. In this work, we present EUROPA, a dataset for multilingual keyphrase generation in the legal domain. It is derived from legal judgments from the Court of Justice of the European Union (EU), and contains instances in all 24 EU official languages. We run multilingual models on our corpus and analyze the results, showing room for improvement on a domain-specific multilingual corpus such as the one we present.

## 1 Introduction

Keyphrases are short phrases that describe a text, and have been shown to be useful for many applications, from document indexing (Medelyan and Witten, 2006) to opinion mining (Berend, 2011). While heavily researched for the STEM (science, technology, engineering, and mathematics) domains, there has been little investigation on its application to law. This is surprising, as keywords can reduce the workload of legal experts by allowing them to get the gist of lengthy documents (Mandal et al., 2017; Sakiyama et al., 2023; Cérat et al., 2023).

The scope of this work is to assess to what extent keyphrases can be automatically generated in the legal domain. Our contributions are the following: (1) We collected and curated EUROPA, the first open benchmark for legal KPG (keyphrase generation), spanning 24 languages and extracted from real-world European judgments.[1] (2) We provide in-depth analysis of our corpus, highlighting its differences compared to other existing corpora. (3) We report performances achieved by multilingual generative models on this benchmark and point out areas where performances could be improved.

## 2 Related Work

While English KPG has been studied fairly extensively in technical and academic domains, our work bridges two less common fields : legal KPG and multilingual KPG. In this section, we first review the literature surrounding keyphrase extraction (KPE) and generation before covering legal KPE and KPG. For a more complete survey on KPG, we refer to Xie et al. (2023).

### 2.1 Keyphrase Extraction and Generation

Keyphrases may be either **present** in the input (could be retrieved with purely **extractive** methods) or **absent** from it (needing **abstractive** methods to generate them).

Many extractive methods have been suggested and used over the years, either using unsupervised heuristic rules and pre-trained machine learning (ML) models to detect and extract keyphrases (Liu et al., 2010; Gollapalli and Caragea, 2014; Meng et al., 2017; Yu and Ng, 2018) or using ML/deep learning (DL) classifiers trained on labelled data to learn to identify them (Yang et al., 2018; Chen et al., 2018; Wang et al., 2018; Basaldella et al., 2018; Alzaidy et al., 2019; Sun et al., 2019; Mu et al., 2020; Sahrawat et al., 2020). Extractive methods are however incomplete as, for most domains, a large portion of the target keyphrases are absent from the document, hence an increased interest in generative models in more recent works.

While research on KPG initially used seq-to-seq models in the form of RNNs (Meng et al., 2017, 2019, 2021; Yuan et al., 2020a), modern KPG uses mostly fine-tuning of pre-trained transformer models (Vaswani et al., 2017), such as BART (Lewis et al., 2020), which showed itself to be as efficient as previous, more complex RNN models (Chowdhury et al., 2022). Current English state-of-the-art

[1] The dataset can be downloaded from the Hugging Face hub at https://huggingface.co/datasets/NCube/europa while evaluation scripts and model outputs are available at https://github.com/rali-udem/europa.

*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12718–12736
August 11-16, 2024 ©2024 Association for Computational Linguistics

results are currently obtained by using BART models which are further pre-trained for scientific KPG (KeyBART by Kulkarni et al. (2022) and SciBART by Wu et al. (2022)). KPG has been conducted on limited domains, such as academic and STEM papers (Hulth, 2003a; Nguyen and Kan, 2007; Kim et al., 2010; Meng et al., 2017; Krapivin et al., 2009) or news (Gallina et al., 2019; Koto et al., 2022), although the need for KPG extends beyond those domains.

Of particular interest is KPG for long documents, which has benefited from a few studies, especially in recent years. Ahmad et al. (2021) proposed a two-step approach of this problem, where they first select salient sentences before using those to generate keyphrases, while Garg et al. (2022) tested multiple ways of including information from the main body (generated summary, citation sentences, random sentences, etc), showing that adding a generated summary was the most efficient strategy. Mahata et al. (2022) proposed a corpus for scientific KPG using the whole document instead of only the title and abstract, which is what is currently used in KPG. They do not however run any baseline on that corpus.

### 2.2 Keyphrase Generation in the Context of Legal Domain

Legal documents are more complex and longer than those commonly used in NLP, as illustrated by the European Court of Human Rights corpus (Chalkidis et al., 2019). This motivates a high demand from legal professionals for automatic digests of such documents (Bendahman et al., 2023), mostly from the angle of legal summarization. This task, more widespread than legal KPG, was first formalized as an extractive task in which the most salient segments from the document are returned as a summary (Farzindar and Lapalme, 2004a,b; Saravanan and Ravindran, 2010; Polsley et al., 2016; Aumiller et al., 2022). Similarly to KPG tasks, abstractive summarization models based on transformer architecture (Vaswani et al., 2017) became more widespread than extractive ones. However, open legal summarization benchmarks such as BigPatent (Sharma et al., 2019), Multi-LexSum (Shen et al., 2022), and EUR-Lex-Sum (Aumiller et al., 2022), are still scarce.

Legal KPG may be considered as a specialized type of summarization but, to the best of our knowledge, no open benchmark is available for such a task. Moreover, most existing works have been focusing on keyphrase extraction (Le et al., 2013; Audich et al., 2016; Mandal et al., 2017; Daojian et al., 2019), ignoring absent keyphrases. The very first legal KPG experiments addressing abstractive keyphrases were conducted by Cérat et al. (2023) and Sakiyama et al. (2023), but within a monolingual setting and with no public dataset release. This is a major issue as generative models are particularly data-greedy. Furthermore, language-wise, most of the open legal datasets are in English (Chalkidis et al., 2023) and, when it comes to existing open multilingual legal benchmarks (Chalkidis et al., 2021; Savelka et al., 2021; Aumiller et al., 2022; Niklaus et al., 2023), none are related to KPG. Overall, our contribution aims at fulfilling these gaps by providing an open multilingual dataset for legal keyphrase generation.

## 3 The Creation of the EUROPA Dataset

Cases in the Court of Justice of the European Union (CJEU), which aims at ensuring the consistent interpretation and application of the EU law across all EU institutions (Court of Justice of the European Union, 2023), can be processed in any of the 24 EU official languages, depending on the Member State involved, making drafting and translation pivotal and complex tasks. Stakeholders and judges communicate in their respective languages and rely on lawyer-linguists for document exchange. Once a judgment is rendered, it is translated into other EU languages to ensure consistency in the judgment text, keyphrases, and their legal implications (Domingues, 2017). Through private correspondence, the CJEU explained that keyphrases, drafted by the Registry and completed by the reporting judge's cabinet or the advocate general, aim at providing concise case descriptions. Their meticulous construction establishes them as high-quality gold references for KPG evaluation.

### 3.1 Data Collection & Processing

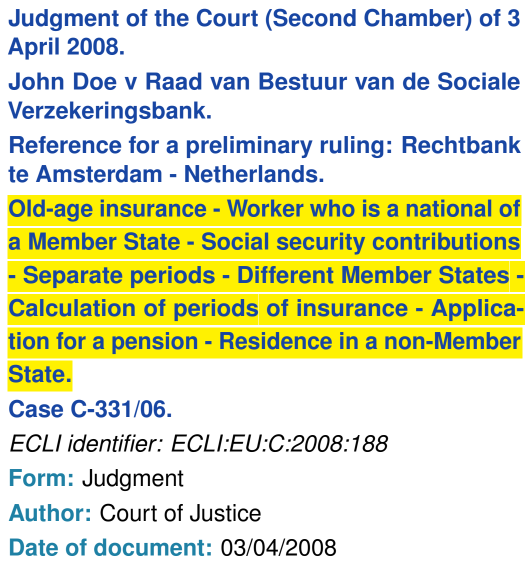
For the sake of clarity, we define a **judgment** as a collection of documents which refer to the same ruling provided in up to 24 languages. Each judgment has a unique CELEX identifier that remains the same for every language version available. Therefore, all language versions with the same CELEX ID are semantically and legally equivalent, and can be considered as parallel. Within a judgment, each language version is named an **instance**, a pair

| | |
|---|---|
| **Input text (fr)** | La demande de décision préjudicielle porte sur l'interprétation de l'article 48 du règlement (CEE) n° 1408/71 du Conseil, du 14 juin 1971, relatif à l'application des régimes de sécurité sociale aux travailleurs salariés, aux travailleurs non salariés et aux membres de leur famille qui se déplacent à l'intérieur de la Communauté (JO L 149, p. 2) [...] |
| **KPs (fr)** | Assurance vieillesse – Travailleur ressortissant d'un État membre – Cotisations sociales – Périodes différentes – États membres différents – Calcul des périodes d'assurance – Demande de pension – Résidence dans un État tiers |
| **Input text (de)** | Das Vorabentscheidungsersuchen betrifft die Auslegung von Art. 48 der Verordnung (EWG) Nr. 1408/71 des Rates vom 14. Juni 1971 zur Anwendung der Systeme der sozialen Sicherheit auf Arbeitnehmer und Selbständige sowie deren Familienangehörige, die innerhalb der Gemeinschaft zu- und abwandern (ABl. L 149, S. 2) [...] |
| **KPs (de)** | Altersversicherung – Arbeitnehmer mit der Staatsangehörigkeit eines Mitgliedstaats – Sozialbeiträge – Unterschiedliche Zeiten – Unterschiedliche Mitgliedstaaten – Berechnung der Versicherungszeiten – Rentenantrag – Wohnort in einem Drittland |
| **Input text (en)** | This reference for a preliminary ruling concerns the interpretation of Article 48 of Council Regulation (EEC) No 1408/71 of 14 June 1971 on the application of social security schemes to employed persons, to self-employed persons and to members of their families moving within the Community (OJ, English Special Edition 1971 (II), p. 416) [...] |
| **KPs (en)** | Old-age insurance – Worker who is a national of a Member State – Social security contributions – Separate periods – Different Member States – Calculation of periods of insurance – Application for a pension – Residence in a non-Member State |

Table 1: Example of a judgment available in 22 languages with the corresponding input text and target keyphrases (French, German and English pairs are shown above). During the KPG task, keyphrases are separated by semicolons instead of dashes to ensure consistency with other KPG datasets.

comprising the target keyphrases and the input text from the judgment, with both being in the same language. For example, the judgment shown in Table 1 has 22 instances spanning 22 languages.

In May 2023, we performed a first pass on the EUR-Lex database[2], the main legal EU online database (Bernet and Berteloot, 2006), scraping 304 426 query results corresponding to 19 319 judgments released by the CJEU. These query results (each being a potential instance for our corpus) consist in small snippets (as the one in Figure 1) containing the multi-line title and meta information of the case including the document identifier, but not the judgment text. Then, we made a second pass in order to scrape the plain text of the judgment using the document identifier collected during the first pass. The plain text is available as PDF and/or HTML, but we used the latter for convenience. A total of approximately two weeks was required for collecting the query results snippets and the judgments' HTML files in all 24 languages.



**Judgment of the Court (Second Chamber) of 3 April 2008.**

**John Doe v Raad van Bestuur van de Sociale Verzekeringsbank.**

**Reference for a preliminary ruling: Rechtbank te Amsterdam - Netherlands.**

**Old-age insurance - Worker who is a national of a Member State - Social security contributions - Separate periods - Different Member States - Calculation of periods of insurance - Application for a pension - Residence in a non-Member State.**

**Case C-331/06.**

*ECLI identifier: ECLI:EU:C:2008:188*

**Form:** Judgment

**Author:** Court of Justice

**Date of document:** 03/04/2008

Figure 1: Query result with multi-line title of a case containing target keyphrases in English (highlighted). The case is the same as the one from Table 1. Non-highlighted text is excluded from targets as it has low value from KPG standpoint.

[2] https://eur-lex.europa.eu

An overview of the corpus collection and curation process is presented in the following paragraphs (more details in Appendix E).

The structure of a case HTML file generally consists of a mix of keyphrases and meta-information at the top of the document followed by paragraphs that will be merged into the input text. Separating the top of the document from the paragraphs is crucial in order to ensure that the input text is not contaminated by target keyphrases. However, identifying keyphrases from the judgment plain text was infeasible as they are surrounded by quotation marks, symbols, and HTML tags that vary across languages and time. Therefore, the keyphrases were obtained from the small snippet multi-line title as the one in Figure 1.

Still, in such snippet, the keyphrases are mixed with meta information about the case which we need to get rid of as they are redundant and of low value from a KPG standpoint (e.g. sequences such as *"Reference for a preliminary ruling"* followed by the referring court and the country are so frequent that they would bring noise during training). Using the `BeautifulSoup` library[3], language-specific regular expressions and domain-aware engineered heuristics, we filtered out the sequence of keyphrases that had the highest number of meaningful phrases. An additional sanity check was also performed to ensure that keyphrases are properly split (e.g. missing whitespace beside a phrase delimiter) and that the number of phrases remains consistent across languages for each judgment. Our approach thus ensures the quality of the target keyphrases that will be used in the KPG task during training and evaluation. For the sake of evaluation consistency with the existing KPG literature, keyphrases are lowercased and separated by semi-colons.

Another critical part of corpus preparation is the input text from the judgment HTML files. These raw documents are delicate to process as they begin with a mix of meta-information (e.g. stakeholders identities) and target keyphrases (which must be excluded from input text), followed by numbered paragraphs. Moreover, since the HTML tags are inconsistent and vary across years and languages, `BeautifulSoup` cannot be used to extract the input text from the judgment. Therefore, we manually designed several language-specific regular expressions to reliably split the documents by matching delimiters in all languages (e.g. *"Judgment"*, *"Sentencia"*, *"Urteil"*). By doing so, only the paragraphs of the judgment are retained as a single input string, while the section containing superficial meta-information and target keyphrases is taken away, thus preventing any data leakage. This was confirmed by a manual examination of 100 random instances, equally distributed across the 24 languages.

After removing instances with empty input or target texts, our final corpus is composed of 17 833 judgments, in 16 languages on average and spanning cases from 1957 to 2023. This amounts to a total of 284 957 instances (input/keyphrases pairs). Expectedly, these instances are unevenly distributed across all 24 languages, with languages from older EU Member States being more represented in the dataset. For instance, French (the most represented language) amounts to 17 461 instances (6.13% of all instances) while Croatian and Irish (a significant outlier) contain 5153 (1.81%) and 92 (0.03%) instances, respectively.

| **Split Type** | **Present** | | **Absent** | |
|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| Random | 30.4 | 41.5 | 15.3 | 18.1 |
| Temporal | 21.7 | 27.4 | 5.6 | 7.0 |

Table 2: Performance for mBART50 with chronological and random splits (weighted average scores).

### 3.2 Chronological Split

It is a common practice to randomly split data in a NLP task (Gorman and Bedrick, 2019). However, training and evaluating a model on data from overlapping time periods with similar distributions fails to assess the actual model's temporal generalization (Lazaridou et al., 2021). This is why legal NLP generally uses a chronological split of documents instead of combining random shuffle with random split (Chalkidis et al., 2019; Medvedeva et al., 2021; Chalkidis et al., 2021). In our case, we tried both splits with a mBART50 model. On the test set, the performance in terms of $F1@M$ for present keyphrases differs by 14.1 percentage points (11.1 for absent keyphrases) in favour of random split. Results in Table 2 confirm that random split, by ignoring real-world temporal concept drifts, tends to overestimate true performance. This is consistent with Søgaard et al. (2021); Mu et al. (2023). Therefore, we choose a chronological split

[3] https://pypi.org/project/beautifulsoup4

| **Benchmark** | **# inst.** | **Avg. input length** | **Avg. num. KPs** | **% absent KPs** | **Domain** | **# lang.** |
|---|---|---|---|---|---|---|
| Inspec (Hulth, 2003b) | 2000 | 116 | 9.6 | 30.0 | Comp. Sc. | 1 (en) |
| Krapivin (Krapivin et al., 2009) | 2304 | 7696 | 5.3 | 23.5 | Comp. Sc. | 1 (en) |
| SemEval (Kim et al., 2010) | 244 | **7795** | 15.4 | 18.3 | Comp. Sc. | 1 (en) |
| KP20k (Meng et al., 2017) | **554k** | 146 | 5.3 | 48.7 | Comp. Sc. | 1 (en) |
| NUS (Nguyen and Kan, 2007) | 211 | 6938 | 11.7 | 17.6 | Science | 1 (en) |
| pak2018 (Campos et al., 2020) | 50 | 96 | 3.6 | **84.2** | Science | 1 (pl) |
| 110-PT-BN-KP (Marujo et al., 2011) | 110 | 301 | **24.4** | 1.3 | News | 1 (pt) |
| WikiNews (Bougouin et al., 2013) | 100 | 268 | 9.6 | 4.8 | News | 1 (fr) |
| KPTimes (Gallina et al., 2019) | 280k | 774 | 5.0 | 55.0 | News | 1 (en) |
| Papyrus (Piedboeuf and Langlais, 2022) | 30k | 307 | 7.2 | 37.2 | Academic | 18 |
| Europa (ours) | 285k | 5220 | 7.6 | 52.6 | Legal | **24** |

Table 3: Comparative table among different open KPG benchmarks. **Avg. input length** refers to the average number of tokens split by whitespaces. Top values are in bold font, second top values are underlined.

for a proper assessment: The training set covers judgments from 1957 to 2010 (131 076 instances), the validation those from 2011 to 2015 (63 373 instances), and the test set the ones from 2016 to 2023 (90 508 instances).[4] Full details about documents distribution across these splits are shown in Appendix F.

## 4 Dataset Analysis

As shown in Table 3, compared to previous works and KPG benchmarks, Europa covers more languages and is highly positioned in matters of number of instances, ratio of absent keyphrases, and average input length.

The distribution of keyphrases in our dataset varies across languages due to the fact that the most recent judgments tend to have a higher number of keyphrases attached to them (the average number of keyphrases per language can be found in Appendix F). Consequently, instances in languages from the most recent Member States tend to be biased towards having more keyphrases.

Another consequence of the temporal evolution of the average number of keyphrases coupled with the chronological split, is that Europa's validation and test sets contain more keyphrases per instance on average (8.3 and 10.5 respectively) compared to its training set (5.4 keyphrases on average). This creates a discrepancy between the sets that practitioners should be aware of. However, we advocate that a higher number of keyphrases in the test set of Europa will be beneficial to the evaluation of the competing models by providing a larger number of potential gold keyphrases, and by assessing the capacity of models to generalize across time when target keyphrases follow different patterns with respect to the past.

## 5 Models

As the majority of Europa's target keyphrases are absent from input documents, we focus on generative models, as extractive ones such as YAKE Campos et al. (2020) are ill-suited here. Recent state-of-the-art models in English KPG rely heavily on pretrained models, such as KeyBART (Kulkarni et al., 2022) or SciBART (Beltagy et al., 2019), which are pre-trained and fine-tuned on a massive corpus of English scientific documents. However, such models are not available in a multilingual format, nor are they specific to the legal domain. We therefore choose to use mBART50 (Tang et al., 2020), a variant of mBART (Liu et al., 2020) with support for 50 languages instead of 25. We also test mT5 (Xue et al., 2021), which covers 101 languages.

Most models, including mBART50, have a maximum input sequence length typically set to 1024 tokens. While mT5's input length can be set arbitrarily, it was originally pretrained with a context of 1024. The main caveat is mT5's memory complex-

[4] The temporal split is available at https://huggingface.co/datasets/NCube/europa and a random split version for those who wish it can be found at https://huggingface.co/datasets/NCube/europa-random-split

| | Weighted Average | | | | | | | Unweighted Average | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | F1 Present | | | F1 Absent | | | MAP | F1 Present | | | F1 Absent | | | MAP |
| | @5 | @10 | @M | @5 | @10 | @M | @50 | @5 | @10 | @M | @5 | @10 | @M | @50 |
| YAKE | 2.0 | 2.2 | 2.2 | 0.0 | 0.0 | 0.0 | 0.2 | 1.9 | 2.1 | 2.1 | 0.0 | 0.0 | 0.0 | 0.2 |
| mT5-small | 11.6 | 7.7 | 17.4 | 2.6 | 1.7 | 4.1 | 4.3 | 11.1 | 7.4 | 16.6 | 2.5 | 1.6 | 3.9 | 4.1 |
| mT5-base | 13.2 | 8.8 | 19.5 | 3.4 | 2.3 | 5.4 | 5.5 | 12.7 | 8.5 | 18.8 | 3.3 | 2.2 | 5.2 | 5.3 |
| mT5-large[a,b] | 13.2 | 8.8 | 19.5 | 3.4 | 2.3 | 5.5 | 5.6 | 12.6 | 8.4 | 18.6 | 3.3 | 2.2 | 5.2 | 5.4 |
| mBART50 | 21.7 | 14.6 | 27.4 | 5.6 | 3.8 | 7.0 | 11.4 | 20.8 | 14.0 | 26.3 | 5.3 | 3.6 | 6.7 | 10.9 |
| mBART50-8k[a] | **23.9** | **16.3** | **29.3** | **5.8** | **3.9** | **7.4** | **12.3** | **23.0** | **15.6** | **28.2** | **5.6** | **3.8** | **7.1** | **11.8** |

Table 4: Weighted and Unweighted average F1@$k$ scores over all languages (%). MAP refers to MAP@50 for all keyphrases combined. Detailed results per language are shown in Appendix G. Highest and second highest scores are in bold font and underlined, respectively. Each model is run once. [a] Due to greater number of parameters and higher training costs, these models were trained during 5 epochs. [b] With a 6th training epoch, mT5-large outperforms mT5-base for some metrics by at most 0.2 percentage points.

ity that increases quadratically with input length, hence a prohibitive computing cost. While there exists some models whose maximum input length reaches up to 16k tokens, such as Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020) or LongT5 (Guo et al., 2022), these remain computationally expensive to run with our current resources and are only suitable for English data. Therefore, similarly to Cérat et al. (2023), we implemented a mBART variant with LSG attention (Condevaux and Harispe, 2023) such that the maximum input length reaches 8192 tokens instead of 1024. Doing so makes the memory complexity increase linearly with respect to the input length, instead of quadratically as it is the case with a traditional attention mechanism. All models are trained with early stopping and a maximum epoch number of 10, except mBART50-8k and mT5-large with only 5 epochs, as their training is more time-consuming. Details about the hyperparameters and training process are provided in Appendices A and D.

### 5.1 Evaluation Protocol

As has become common practice in recent works on KPG (Meng et al., 2017; Garg et al., 2023; Shen and Le, 2023; Chen and Iwaihara, 2023), model evaluation relies on two F1 measures: F1@$k$ ($k = \{5, 10\}$) and F1@$M$, calculated separately for present and absent keyphrases. F1@$M$ is computed using the entirety of the generated keyphrases while F1@$k$ is computed using the best $k$ generated keyphrases (by truncating if necessary). In both cases, the number of target keyphrases remain untouched. F1@$k$ is calculated using only the top $k$ best scoring keyphrases among the model's predictions, hence an upper boundary below 1 whenever the number of target keyphrases exceeds $k$, which occurs in most of Europa's instances. This is why F1@$M$ tends to better reflect the model performance as all candidate keyphrases are taken into account without being truncated. For more details about these metrics, we refer to Yuan et al. (2020b). Following the literature, target and predicted keyphrases are lowercased and stemmed (e.g. Meng et al. (2017) applied Porter Stemmer for English) before conducting an exact match. Stemming is a critical step without which a candidate keyphrase could be errouneously considered wrong because of the morphological nature of the language. That is why we applied stemming for all languages for treating them as fairly as possible. For most of them, the Snowball stemmer (Porter, 2001) was used. In addition to F1 measures, we also computed MAP@50 (Mean Average Precision).

## 6 Results

$F1@5$, $F1@10$ and $F1@M$ scores for present and absent keyphrases over all languages are reported in Table 4. Since the number of instances varies across languages, we report average scores that are **weighted** and **unweighted**. The former is an average score across all instances without taking language into consideration. Consequently, it can be highly influenced by high-resource languages covering more instances. The latter is an unweighted average among languages' scores. In other words, all languages have equal importance, thus reflecting the ability of a model to perform equally well in high- and low-resource languages.

Unsurprisignly, the YAKE extractive model performs poorly, finding none of the absent keyphrases.

| Language | Present | | Absent | |
|---|---|---|---|---|
| | 1k | 8k | 1k | 8k |
| French | 33.1 | **34.4** | 8.7 | **9.6** |
| German | 33.5 | **35.8** | 4.9 | **5.0** |
| English | 29.5 | **31.7** | 5.0 | **5.1** |
| Italian | 30.5 | **33.3** | **3.7** | 3.6 |
| Dutch | 31.9 | **34.4** | 3.5 | **3.7** |
| Greek | 14.8 | **15.7** | 10.6 | **11.1** |
| Danish | 30.4 | **32.6** | **3.2** | 2.6 |
| Portuguese | 27.6 | **30.1** | 5.5 | **6.5** |
| Spanish | 29.2 | **31.8** | 4.8 | 4.8 |
| Swedish | 33.1 | **33.9** | 4.2 | **5.0** |
| Finnish | 27.5 | **28.9** | 11.6 | **13.1** |
| Lithuanian | **34.9** | 32.9 | 4.2 | **4.7** |
| Estonian | 29.5 | **31.0** | **4.7** | 4.5 |
| Czech | 26.4 | **29.5** | 13.7 | **14.6** |
| Hungarian | 14.3 | **15.3** | 10.4 | **11.3** |
| Latvian | **34.7** | 33.1 | 3.5 | **3.6** |
| Slovene | 25.6 | **29.2** | 11.1 | **11.8** |
| Polish | 26.3 | **29.7** | 11.7 | **13.2** |
| Maltese | 24.7 | **27.0** | **7.7** | 7.1 |
| Slovak | **25.2** | 24.9 | 15.7 | **16.0** |
| Romanian | 30.7 | **33.8** | 5.4 | **5.7** |
| Bulgarian | 29.0 | **29.2** | 3.4 | **3.5** |
| Croatian | 8.6 | **15.5** | **4.1** | 3.8 |
| Irish | 0.0 | **2.2** | 0.0 | 0.0 |

Table 5: Side-by-side F1@M scores comparison per language between mBART50 and mBART50-8k.

With a fixed input length of 1024 tokens, the three mT5 variants dramatically underperform mBART50. This is surprising as mT5 covers twice as many languages than mBART50. Also, as mT5-large has 59% more parameters than mBART50, it would have been expected to outperform the latter, but the results reveal otherwise. When comparing mBART50 with mBART50-8k, the increase in the maximum context length brings significant improvement across all metrics. While an average gain of around 2% (for $F1@M$ over present phrases) may not seem significant, KPG evaluation often greatly underestimates the true performance of the models, due to the difficulty of correctly evaluating whether a keyphrase is correct or not (Wu et al., 2023). As such, the improvement shown by the mBART50-8k model is significant and reflected in the generated keyphrases, thus emphasizing the benefits in enlarging the maximum input length. Still, Table 5 shows that input context enlargement does not have uniform effects across languages. For instance, for present keyphrases, some languages get small gains (Greek, Swedish), and some degrade (Lithuanian, Latvian). For high-resource languages such as English and Italian, performance improves on present phrases, but seems stagnant on absent ones. For low-resource languages, Croatian has the most dramatic improvement on present phrases while Irish gets one score above 0 with mBART50-8k. However, the performance for these languages still lags behind that of moderate-resource ones. This is understandable as these languages have almost no training instances in our temporal split setting, thus revealing the difficulty of conducting KPG for unseen languages.[5]

## 7 Analysis and Discussion

The first striking observation is that mBART models, despite covering less languages compared to mT5s, succeed in outperforming the latter models. One explanation is that mT5 small, base and large variants generated on average significantly fewer phrases per instance (2.1, 2.3 and 2.3, respectively) with respect to mBART50 and mBART50-8k (5.5 and 6.0). Consequently, mT5 models are less likely to achieve high scores. The other noticeable result is that the mBART models even succeed in outperforming mT5s in languages that only mT5s support such as Bulgarian or Greek. This suggests that Conneau et al. (2020)'s corpus used for pretraining mBART50 gave it a capacity to deal with much more languages than stated by Tang et al. (2020).

The improvement from mBART50 to mBART50-8k is significant and emphasizes the importance of larger input length. Less that 14% of all instances fit into a 1024-tokens context window. Around 58% do when that length reaches 8k tokens. Building models that can efficiently generate keyphrases from larger documents is therefore crucial in order to achieve further progress (the length reaches 9160 tokens on average, and 17k at the 90th percentile, with mBART50 tokenizer).[6]

Moreover, KPG models struggle for keyphrases with more tokens (split by whitespaces). In the case of mBART50-8k, a matched target keyphrase contains on average 3.1 tokens, while an unmatched

[5] With a random split dataset, $F1@M$ achieved by mBART50 for Croation/Irish reaches 46.9/20.5 for present keyphrases and 13.9/10.2 for absent ones

[6] We tried a sliding-window-based model that was not conclusive, thus the need to find better ways to capture context.

**Reference**: Appeal [5] – Competition [5] – *Agreements, decisions and concerted practices* [3] – Pharmaceutical products – *Market for antidepressant medicines (citalopram)* – *Settlement agreements concerning process patents concluded between a manufacturer of originator medicines holding those patents and manufacturers of generic medicines* – Article 101 TFEU [1] – Potential competition – Restriction by object – Characterisation – *Calculation of the amount of the fine*
**mT5-small**: Appeal – Competition – Regulation (EC) No 1/2003
**mT5-base**: Appeal – Competition – *Agreements, decisions and concerted practice*
**mT5-large**: Appeal – Competition – Regulation (EC) No 1/2003
**mBART50**: Appeal – Competition – *Agreements, decisions and concerted practices* – Non-contractual liability of the Community – Guidelines on the application of Article 101 TFEU – Principle of proportionality – Obligation to state the reasons on which the decision is based
**mBART50-8k**: Appeal – Competition – Article 101 TFEU – *Agreements, decisions and concerted practices* – Antidepressant medicinal products – Article 23(2)(a) of Regulation (EC) No 1/2003 – Article 23(2)(a) of Regulation (EC) No 1/2003 – Concept of 'restrictions of competition by object' – Reduction of the amount of the fine

Table 6: Example of generated keyphrases with an instance in English. Blue phrases are exact matched targets, purple ones are candidates which could be relevant but are not rewarded by the exact match evaluation approach. We added comments in grey with the number of times some targets were matched. Absent keyphrases are in *italics*.

target phrase contains 5.7 tokens. Unmatching generated keyphrases contain 7.7 tokens on average. Upon manual inspection of generated keyphrases, most models succeed in generating simple phrases made of up to 3 terms, but they indeed struggle for longer noun phrases that refer to very specific or technical concepts such as *"Market for antidepressant medicines (citalopram)"* in Table 6. Also, although some candidate keyphrases are still relevant from a reader's standpoint, they are penalized by the exact matching evaluation approach, although stemming is applied. For instance, the output *"Concept of 'restrictions of competition by object'"* could be a decent generation for the target *"Restriction by object"* despite the lack of matching.

In order to mitigate this issue, we evaluated the models again with a semantic matching metric (Wu et al., 2023) allowing us to compare predictions and targets without having to apply any sort of post-processing or stemming (the tool is however only available for English). Results in Table 7 have a correlation of 0.99 with F1@$M$ scores obtained for English for both present and absent keyphrases. This seems consistent with the ranking among models observed previously in Table 4: mBART50-based models outperform mT5-based ones, and mBART50-8k is the front-runner.

| **Model** | **Present** | **Absent** |
|---|---|---|
| YAKE | 0.335 | 0.155 |
| mT5-small | 0.399 | 0.268 |
| mT5-base | 0.422 | 0.283 |
| mT5-large | 0.436 | 0.270 |
| mBART50 | 0.519 | 0.442 |
| mBART50-8k | **0.548** | **0.479** |

Table 7: Semantic matching scores for present and absent keyphrases in English.

## 8 Conclusion

In this paper, we present EUROPA, a novel and open multilingual keyphrase generation dataset in the legal domain. We believe this dataset can help alleviate two current shortcomings of the keyphrase generation task: the lack of data in domains other than STEM; and the lack of multilingual non-English datasets. Our dataset is available at `https://huggingface.co/datasets/NCube/europa`. We furthermore provide an analysis of EUROPA with key statistics, thus giving insights on the particularities of our dataset. Finally, we run multiple models in various settings in order to give an initial point of comparison for future works on EUROPA. Our corpus also highlights the need to efficiently capture larger input context, and will be a suitable testbed for models designed to do so.

## Limitations

**Low Resource Languages:** The choice of a chronological split, though enabling a realistic KPG performance assessment with quasi-equal amount of test instances for each language (except Irish), results in dramatic differences in available training data across languages. This is particularly true for the latest official EU languages, such as Croatian and Irish. One possible approach for mitigating this issue in future works would be to complement the training data with other legal documents released by other EU institutions. A random split version of our dataset is available for those who would like an identical distribution of languages across training, validation and test sets (`https://huggingface.co/datasets/NCube/europa-random-split`).

**Computing Cost Scalability:** The architecture of most of the models used here is derived from experiments in which the text input length rarely exceeds 1024. This is a major caveat when deploying such models on real-world data with significantly longer documents, such as the legal documents in our dataset which often exceed several thousand tokens. Moreover, dealing with documents whose language is not supported by the tokenizer mechanically increases the number of tokens. This is due to the fact that running and training transformer models with higher maximum input sequences is costly both in terms of time and GPU memory. The computing burden was already emphasized by Aumiller et al. (2022) of legal summarization for EU documents and Sakiyama et al. (2023) for legal KPG in Portuguese.

## Ethics Statement

**Copyright:** The Publications Office of the European Union gave us the written confirmation that cases of the Court of Justice of the European Union could be used for commercial and non-commercial purposes, as stated in the European Commission decision released on 12 December 2011 (European Commission, 2011). Our models are implemented with Wolf et al. (2020)'s `transformers` library licensed under Apache 2.0 while evaluation is performed with Meng et al. (2017)'s tools available under the MIT license. Therefore, one is granted permission to modify and distribute the licensed material.

**Personal Data Protection:** According to the CJEU personal data protection policy, the Court anonymizes personal information upon request of a party, referring court/tribunal or upon its own volition. In that case, anonymity is granted throughout the entire procedure according to article 95 of Rules of Procedure of the Court of Justice. It must be emphasized that anonymity must be requested at the earliest stage of the proceedings. Once the decision is drafted and released, there is no a posteriori opt-out option due to the transparency principle with which the CJEU must comply. Moreover, we have a written confirmation from the Publications Office of the EU that we are allowed to share and redistribute our corpus made from documents publicly available on EU platforms.

## Acknowledgements

We would like to thank the European Union for granting us access to EUR-Lex data (`http://eur-lex.europa.eu`, © European Union, 1998-2023). We are also grateful to Benjamin Cérat and Di Wu for their advice on some coding implementations.

## References

Wasi Ahmad, Xiao Bai, Soomin Lee, and Kai-Wei Chang. 2021. Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. 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 1389–1404, Online. Association for Computational Linguistics.

Rabah Alzaidy, Cornelia Caragea, and C Lee Giles. 2019. Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In *The world wide web conference*, pages 2551–2557.

Dhiren A Audich, Rozita Dara, and Blair Nonnecke. 2016. Extracting keyword and keyphrase from online privacy policies. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, pages 127–132. IEEE.

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Basaldella, Elisa Antolli, Giuseppe Serra, and Carlo Tasso. 2018. Bidirectional lstm recurrent neural network for keyphrase extraction. In *Digital Libraries and Multimedia Archives: 14th Italian Research Conference on Digital Libraries, IRCDL 2018, Udine, Italy, January 25-26, 2018, Proceedings 14*, pages 180–187. Springer.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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

Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, and Mokhtar Boumedyen Billami. 2023. Quelles évolutions sur cette loi? entre abstraction et hallucination dans le domaine du résumé de textes juridiques. In *18ème Conférence en Recherche d'Information et Applications (CORIA 2023)*, pages 18–36. ATALA.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, 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.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Bin Chen and Mizuho Iwaihara. 2023. Enhancing keyphrase generation by bart finetuning with splitting and shuffling. In *Pacific Rim International Conference on Artificial Intelligence*, pages 305–310. Springer.

Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. Keyphrase generation with correlation constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium. Association for Computational Linguistics.

Md Faisal Mahbub Chowdhury, Gaetano Rossiello, Michael Glass, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2022. Applying a generic sequence-to-sequence model for simple and effective keyphrase generation. *arXiv preprint arXiv:2201.05302*.

Charles Condevaux and Sébastien Harispe. 2023. Lsg attention: Extrapolation of pretrained transformers to long sequences. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part I*, pages 443–454. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Court of Justice of the European Union. 2023. Annual report 2022.

Benjamin Cérat, Olivier Salaün, Noreddine Ben Jillali, Marc-André Morissette, Isabela Pocovnicu, Emma Elliot, and François Harvey. 2023. LexKey: A Keyword Generator for Legal Documents. In *Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal*.

Zeng Daojian, Tong Guowei, Dai Yuan, Li Feng, Han Bing, and Xie Songxian. 2019. Keyphrase extraction for legal questions based on a sequence to sequence model. *Journal of Tsinghua University (Science and Technology)*, 59(4):256–261.

Joana Sousa Domingues. 2017. The multilingual jurisprudence of the court of justice and the idea of uniformity in european union law. *UNIO–EU Law Journal*, 3(2):125–138.

European Commission. 2011. Commission Decision of 12 December 2011 on the reuse of Commission documents.

Atefeh Farzindar and Guy Lapalme. 2004a. Legal text summarization by exploration of the thematic structure and argumentative roles. In *Text Summarization*

*Branches Out*, pages 27–34, Barcelona, Spain. Association for Computational Linguistics.

Atefeh Farzindar and Guy Lapalme. 2004b. Letsum, an automatic legal text summarizing. In *Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference*, volume 120, page 11. IOS Press.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2019. Kptimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135.

Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. 2022. Keyphrase generation beyond the boundaries of title and abstract. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5809–5821, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. 2023. Data augmentation for low-resource keyphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8442–8455, Toronto, Canada. Association for Computational Linguistics.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Anette Hulth. 2003a. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

Anette Hulth. 2003b. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Lipkey: A large-scale news dataset for absent keyphrases generation and abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3427–3437.

Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction.

Mayank Kulkarni, Debanjan Mahata, Ravneet Arora, and Rajarshi Bhowmik. 2022. Learning rich representation of keyphrases from text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 891–906, Seattle, United States. Association for Computational Linguistics.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu. 2013. Unsupervised keyword extraction for japanese legal documents. In *JURIX*, pages 97–106.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376, Cambridge, MA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Debanjan Mahata, Navneet Agarwal, Dibya Gautam, Amardeep Kumar, Swapnil Parekh, Yaman Kumar Singla, Anish Acharya, and Rajiv Ratn Shah. 2022. Ldkp: A dataset for identifying keyphrases from long scientific documents. *arXiv preprint arXiv:2203.15349*.

Arpan Mandal, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2017. Automatic catchphrase identification from legal court case documents. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2187–2190.

Luis Marujo, Márcio Viveiros, and João Paulo da Silva Neto. 2011. Keyphrase cloud generation of broadcast news. *12th Annual Conference of the International Speech Communication Association 2011 (INTERSPEECH 2011)*, pages 2404–2407.

Olena Medelyan and Ian H Witten. 2006. Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*.

Rui Meng, Xingdi Yuan, Tong Wang, Peter Brusilovsky, Adam Trischler, and Daqing He. 2019. Does order matter? an empirical study on generating multiple keyphrases as a sequence. *arXiv preprint arXiv:1909.03590*.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–5007, Online. Association for Computational Linguistics.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Funan Mu, Zhenting Yu, LiFeng Wang, Yequan Wang, Qingyu Yin, Yibo Sun, Liqun Liu, Teng Ma, Jing Tang, and Xing Zhou. 2020. Keyphrase extraction with span-based feature representations. *arXiv preprint arXiv:2002.05407*.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743, Dubrovnik, Croatia. Association for Computational Linguistics.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *International conference on Asian digital libraries*, pages 317–326. Springer.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Frédéric Piedboeuf and Philippe Langlais. 2022. A new dataset for multilingual keyphrase generation. *Advances in Neural Information Processing Systems*, 35:38046–38059.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.

M.F. Porter. 2001. Snowball: A language for stemming algorithms.

Dhruva Sahrawat, Debanjan Mahata, Haimin Zhang, Mayank Kulkarni, Agniv Sharma, Rakesh Gosangi, Amanda Stent, Yaman Kumar, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Keyphrase extraction as sequence labeling using contextualized embeddings. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 328–335. Springer.

Kenzo Sakiyama, Rodrigo Nogueira, and Roseli A. F. Romero. 2023. Automated keyphrase generation for brazilian legal information retrieval. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

M Saravanan and Balaraman Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18(1):45–76.

Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S Alexander, Jayla C Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, et al. 2021. Lex rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 129–138.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Lingyun Shen and Xiaoqiu Le. 2023. An enhanced method on transformer-based model for one2seq keyphrase generation. *Electronics*, 12(13):2968.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Zhiqing Sun, Jian Tang, Pan Du, Zhi-Hong Deng, and Jian-Yun Nie. 2019. Divgraphpointer: A graph pointer network for extracting diverse keyphrases. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 755–764.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yanan Wang, Qi Liu, Chuan Qin, Tong Xu, Yijun Wang, Enhong Chen, and Hui Xiong. 2018. Exploiting topic-based adversarial neural network for cross-domain keyphrase extraction. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 597–606. IEEE.

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.

Di Wu, Wasi Uddin Ahmad, and Kai-Wei Chang. 2022. Pre-trained language models for keyphrase generation: A thorough empirical study. *arXiv preprint arXiv:2212.10233*.

Di Wu, Da Yin, and Kai-Wei Chang. 2023. Kpeval: Towards fine-grained semantic-based evaluation of keyphrase extraction and generation systems. *arXiv preprint arXiv:2303.15422*.

Binbin Xie, Jia Song, Liangying Shao, Suhang Wu, Xiangpeng Wei, Baosong Yang, Huan Lin, Jun Xie, and Jinsong Su. 2023. From statistical methods to deep learning, automatic keyphrase prediction: A survey. *Information Processing & Management*, 60(4):103382.

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.

Min Yang, Yuzhi Liang, Wei Zhao, Wei Xu, Jia Zhu, and Qiang Qu. 2018. Task-oriented keyphrase extraction from social media. *Multimedia Tools and Applications*, 77:3171–3187.

Yang Yu and Vincent Ng. 2018. Wikirank: Improving keyphrase extraction based on background knowledge. *arXiv preprint arXiv:1803.09000*.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2020a. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2020b. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

## A Hyperparameters

While the batch size was set at 16 (with the help of gradient batch accumulation when needed), a grid search was performed for the learning rate with the following values: $1 \times 10^{-6}$, $5 \times 10^{-6}$, $1 \times 10^{-5}$, $5 \times 10^{-5}$, $1 \times 10^{-4}$. For both mBART50 models, we settled on a learning rate of $1 \times 10^{-5}$ which we coupled with the `AdamW` (Loshchilov and Hutter, 2018) optimizer of `Pytorch`. The same applies to mT5 variants but with a learning rate of $1 \times 10^{-4}$. All other optimizer hyperparameters are left at default values. Once the learning rate was determined, we also added 2500 warming steps and a scheduler with a linear decay. The objective function was cross-entropy loss. For mBART models, we applied mixed precision (FP16) while brainfloat16 (BF16) was used for mT5 models. The maximum number of epochs was set to 10 and the model achieving the lowest loss on the validation set is used for the test set. Patience was set to 5 epochs. In the case of large models (mBART-8k, mT5-large), the maximum number of epochs was set to 5 due to much longer training epochs. For generation, we used a beam search of 10 and return the best sequence, with a maximum output length of 256 tokens. The hardware consisted in a RTX4090 and a couple of A5000 graphical cards, each with approximately 24 GB of VRAM. Inference only required a single card for all models, including those with the highest number of parameters. Overall, the training for either mBART-8k or mT5-large was performed on 2 GPUs for up to 62 hours overall. For other models, the training lasted for up to 28h on a single GPU (more details in Table 10). The inference on the test set took up to 34 hours, depending on the model. Given the delay in obtaining keyphrases during inference, the results are based on a single run.

## B Languages Support per Model

mBART50 by Tang et al. (2020) and mT5 by Xue et al. (2021), though presented as multilingual models, do not cover the same languages and both support only partially the official EU languages, as shown in Table 8. Although mT5 (its tokenizer remains the same for all variant sizes) supports twice as many languages than mBART, the vocabulary of both tokenizers is roughly of the same size: 250 112 and 250 054 tokens, respectively. Such large vocabulary size is a computing burden during inference time, especially during decoding.

By design, mBART50 tokenizer adds to input and output sequences a token specifying source and target languages (e.g. [en_XX] for English). As not all EU languages are not covered by mBART50, we manually added to the tokenizer special tokens for all unsupported languages (e.g. [bg_BG] for Bulgarian). Following mBART50 tokenization process, a language token is added in both input and output sequences. Unlike translation tasks, our multilingual KPG task is such that, for each instance, the input document and the target keyphrases are both in the same language. Therefore, the language token remains the same in the input and output within the same instance.

When it comes to mT5, such languages prefixes are not required. However, applying mBART50 language tokens to mT5-small brought around 1 percentage point improvement across F1@$k$ metrics, thus motivating us to deploy mBART50 language prefixes to the rest of mT5 models.

| **Model** | mBART50 | mT5 |
|---|---|---|
| **Total lang.** | 52 | 101 |
| **Unsupported EU lang.** | bg, da, el, ga, hu, mt, sk | hr |

Table 8: Language Coverage by Multilingual Models

## C Number of Parameters per Model

| **Model** | **# param. (millions)** |
|---|---|
| google/mt5-small | 172 |
| google/mt5-base | 390 |
| google/mt5-large | 973 |
| facebook/mbart-large-50 | 611 |
| mBART-8k | 626 |

Table 9: Number of parameters (in millions) for each model. Each model name corresponds to its identifier on Wolf et al. (2020)'s Hugging Face platform, expect mBART-8k which we produced by applying LSG attention (Condevaux and Harispe, 2023) to mBART.

## D Training and Inference Times

| **Model** | **Num. of training epochs** | **Total training time** | **Total inference time** |
|---|---|---|---|
| mT5-small | 10 | 11h | 7h |
| mT5-base | 10 | 28h | 9h |
| mT5-large* | 5 | 53h | 16h |
| mBART50 | 10 | 25h | 31h |
| mBART50-8k* | 5 | 61h | 34h |

Table 10: Total training and inference times for each model. Total training time corresponds to the duration of all training epochs, including computation of loss on the validation set. Total inference time refers to the duration required for generating keyphrases over the entire test set. All measurements correspond to a single run. *Due to greater number of parameters and higher training costs, these models were trained during 5 epochs.

# E Details About the Data Preprocessing

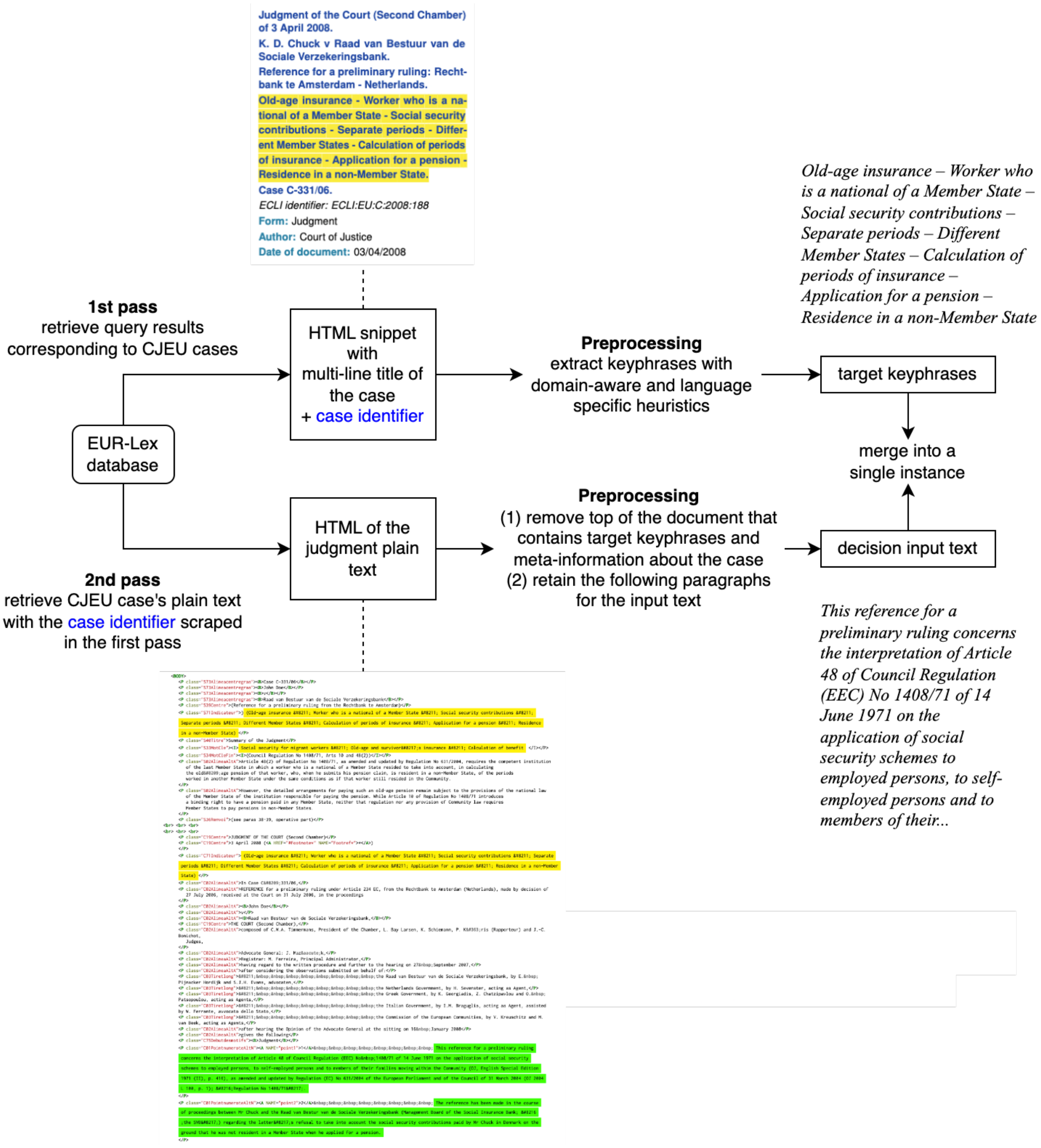

Judgment of the Court (Second Chamber) of 3 April 2008.
K. D. Chuck v Raad van Bestuur van de Sociale Verzekeringsbank.
Reference for a preliminary ruling: Rechtbank te Amsterdam - Netherlands.
Old-age insurance - Worker who is a national of a Member State - Social security contributions - Separate periods - Different Member States - Calculation of periods of insurance - Application for a pension - Residence in a non-Member State.
Case C-331/06.
ECLI identifier: ECLI:EU:C:2008:188
Form: Judgment
Author: Court of Justice
Date of document: 03/04/2008
1st pass
retrieve query results corresponding to CJEU cases
HTML snippet with multi-line title of the case + case identifier
Preprocessing
extract keyphrases with domain-aware and language specific heuristics
target keyphrases
Old-age insurance – Worker who is a national of a Member State – Social security contributions – Separate periods – Different Member States – Calculation of periods of insurance – Application for a pension – Residence in a non-Member State
EUR-Lex database
merge into a single instance
HTML of the judgment plain text
Preprocessing
(1) remove top of the document that contains target keyphrases and meta-information about the case
(2) retain the following paragraphs for the input text
decision input text
2nd pass
retrieve CJEU case's plain text with the case identifier scraped in the first pass
This reference for a preliminary ruling concerns the interpretation of Article 48 of Council Regulation (EEC) No 1408/71 of 14 June 1971 on the application of social security schemes to employed persons, to self-employed persons and to members of their...

Figure 2: Diagram illustrating the data collection procedure. Target keyphrases and input text are extracted from separate pieces of HTML that are linked by a common case identifier. After preprocessing, they are merged into a single instance. The same process is repeated in every language.

## F Additional Statistics

| Set | # judg. | # inst. | Avg. num. KPs | % present KPs | % absent KPs |
|---|---|---|---|---|---|
| Train (1957-2010) | 10 003 | 131 076 | 5.4 | 45.6 | 54.4 |
| Valid. (2011-2015) | 3 391 | 63 373 | 8.3 | 45.1 | 54.9 |
| Test (2016-2023) | 4 439 | 90 508 | 10.5 | 51.7 | 48.3 |

Table 11: Statistics on the training, validation, and test sets of Europa. **# judg.** and **# inst.** stand for the numbers of judgments (judgment is a ruling released in several languages) and instances (an instance is a pair with input text and target keyphrases both in a single language).

| Language | Lang. ISO code | Official EU lang. since | Train set size | Valid. set size | Test set size | Avg. num. KPs | % absent KPs | Avg. KPs length | Avg. input length |
|---|---|---|---|---|---|---|---|---|---|
| French | fr | 1958 | 9692 | 3338 | 4431 | 7.0 | 51.6 | 5.5 | 5651 |
| German | de | 1958 | 9028 | 2817 | 3962 | 6.9 | 47.8 | 4.6 | 4877 |
| English | en | 1973 | 9047 | 2826 | 3950 | 6.8 | 46.8 | 5.6 | 5539 |
| Italian | it | 1958 | 8978 | 2831 | 3946 | 6.9 | 53.2 | 5.6 | 5402 |
| Dutch | nl | 1958 | 9066 | 2809 | 3875 | 6.8 | 50.5 | 5.0 | 5374 |
| Greek | el | 1981 | 8945 | 2833 | 3917 | 6.9 | 73.9 | 5.3 | 3894 |
| Danish | da | 1973 | 9018 | 2702 | 3886 | 6.8 | 52.2 | 4.8 | 5039 |
| Portuguese | pt | 1986 | 8347 | 2801 | 3913 | 7.1 | 56.0 | 5.8 | 5763 |
| Spanish | es | 1986 | 8462 | 2807 | 3932 | 7.0 | 49.4 | 6.1 | 6154 |
| Swedish | sv | 1995 | 6736 | 2788 | 3903 | 7.5 | 45.8 | 4.8 | 5392 |
| Finnish | fi | 1995 | 6874 | 2806 | 3886 | 7.5 | 56.5 | 4.0 | 4139 |
| Lithuanian | lt | 2004 | 3598 | 2828 | 3908 | 8.5 | 47.1 | 4.8 | 4674 |
| Estonian | et | 2004 | 3621 | 2810 | 3912 | 8.4 | 42.4 | 4.0 | 4363 |
| Czech | cs | 2004 | 3621 | 2804 | 3913 | 8.4 | 54.9 | 5.1 | 5167 |
| Hungarian | hu | 2004 | 3625 | 2805 | 3912 | 8.4 | 55.1 | 4.8 | 5138 |
| Latvian | lv | 2004 | 3627 | 2814 | 3896 | 8.5 | 47.1 | 4.8 | 4825 |
| Slovene | sl | 2004 | 3568 | 2800 | 3863 | 8.4 | 48.7 | 4.8 | 5187 |
| Polish | pl | 2004 | 3526 | 2732 | 3918 | 8.4 | 63.8 | 5.1 | 5385 |
| Maltese | mt | 2004 | 3394 | 2739 | 3905 | 8.5 | 58.1 | 5.1 | 5185 |
| Slovak | sk | 2004 | 3337 | 2803 | 3894 | 8.5 | 56.9 | 5.1 | 5233 |
| Romanian | ro | 2007 | 2484 | 2812 | 3916 | 8.7 | 51.5 | 5.8 | 6085 |
| Bulgarian | bg | 2007 | 2480 | 2822 | 3873 | 8.7 | 47.0 | 5.8 | 5827 |
| Croatian | hr | 2013 | 2 | 1246 | 3905 | 10.1 | 46.0 | 5.2 | 6367 |
| Irish | ga | 2022 | 0 | 0 | 92 | 11.9 | 48.6 | 6.5 | 10101 |
| Overall | | | 131 076 | 63 373 | 90 508 | 7.6 | 52.6 | 5.1 | 5220 |

Table 12: Distribution of documents across languages and splits. High-volume languages are those spoken by the earliest EU Member States (e.g. France and Germany are among the Founding States of the EU). **Avg. KPs length** refers to the average number of words per keyphrase (split at whitespaces). **Avg. input length** refers to the average number of tokens in the input text (split at whitespaces).

## G Scores per Language for each Model

| Model | Bulgarian * | | Croatian † | | Czech | | Danish * | | Dutch | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 1.8 | 1.9 | 2.6 | 2.4 | 3.0 | 2.8 | 0.6 | 1.0 | 2.1 | 2.5 |
| mT5-small | 13.3 | 19.9 | 1.5 | 2.4 | 9.5 | 15.3 | 15.1 | 21.5 | 15.4 | 22.4 |
| mT5-base | 14.7 | 21.5 | 3.9 | 6.1 | 10.9 | 17.2 | 16.8 | 23.9 | 16.3 | 23.5 |
| mT5-large | 14.6 | 21.6 | 3.3 | 5.1 | 11.2 | 17.8 | 16.6 | 23.4 | 16.4 | 23.4 |
| mBART50 | 23.0 | 29.0 | 5.8 | 8.6 | 20.0 | 26.4 | 25.0 | 30.4 | 25.5 | 31.9 |
| mBART50-8k | **23.7** | **29.2** | **11.8** | **15.5** | **23.2** | **29.5** | **27.6** | **32.6** | **28.3** | **34.4** |

| Model | English | | Estonian | | Finnish | | French | | German | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 2.7 | 3.4 | 1.9 | 2.5 | 1.1 | 1.2 | 2.4 | 2.5 | 2.4 | 3.1 |
| mT5-small | 15.0 | 20.5 | 13.2 | 18.4 | 9.3 | 14.8 | 15.4 | 23.0 | 16.7 | 23.8 |
| mT5-base | 16.1 | 21.9 | 15.5 | 21.3 | 12.2 | 19.1 | 16.6 | 24.3 | 17.8 | 24.9 |
| mT5-large | 16.4 | 22.2 | 15.6 | 21.5 | 12.0 | 19.0 | 16.7 | 24.5 | 17.9 | 25.2 |
| mBART50 | 24.9 | 29.5 | 25.5 | 29.5 | 19.1 | 27.5 | 26.6 | 33.1 | 28.7 | 33.5 |
| mBART50-8k | **27.8** | **31.7** | **27.2** | **31.0** | **21.1** | **28.9** | **28.8** | **34.4** | **31.1** | **35.8** |

| Model | Greek * | | Hungarian * | | Irish * | | Italian | | Latvian | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 1.0 | 0.9 | 2.1 | 1.9 | 0.7 | 1.0 | 2.9 | 3.0 | 2.4 | 2.8 |
| mT5-small | 6.3 | 11.0 | 7.4 | 11.3 | 0.2 | 0.2 | 13.8 | 21.0 | 12.7 | 18.7 |
| mT5-base | 8.3 | 14.2 | 9.1 | 13.7 | 1.3 | 1.8 | 14.8 | 21.9 | 15.0 | 21.4 |
| mT5-large | 7.7 | 13.2 | 9.0 | 13.6 | 0.0 | 0.0 | 14.1 | 20.9 | 14.0 | 20.2 |
| mBART50 | 9.3 | 14.8 | 10.0 | 14.3 | 0.0 | 0.0 | 23.8 | 30.5 | 28.4 | 34.7 |
| mBART50-8k | **10.7** | **15.7** | **10.9** | **15.3** | **1.4** | **2.2** | **27.2** | **33.3** | **27.8** | **33.1** |

| Model | Lithuanian | | Maltese * | | Polish | | Portuguese | | Romanian | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 1.6 | 1.7 | 0.3 | 0.4 | 2.3 | 2.2 | 2.5 | 2.6 | 1.7 | 1.7 |
| mT5-small | 12.8 | 18.7 | 10.1 | 15.8 | 8.7 | 14.3 | 7.2 | 11.4 | 13.7 | 20.2 |
| mT5-base | 16.2 | 23.3 | 11.3 | 17.7 | 10.7 | 17.6 | 8.4 | 13.0 | 14.0 | 20.5 |
| mT5-large | 15.2 | 21.8 | 11.4 | 17.7 | 10.9 | 17.9 | 8.5 | 13.2 | 14.7 | 21.6 |
| mBART50 | **28.8** | **34.9** | 18.8 | 24.7 | 19.1 | 26.3 | 21.6 | 27.6 | 24.9 | 30.7 |
| mBART50-8k | 27.9 | 32.9 | **21.5** | **27.0** | **21.6** | **29.7** | **24.6** | **30.1** | **28.2** | **33.8** |

| Model | Slovak * | | Slovene | | Spanish | | Swedish | |
|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 1.6 | 1.7 | 2.3 | 2.6 | 2.3 | 2.5 | 2.0 | 2.9 |
| mT5-small | 10.7 | 17.2 | 9.9 | 14.7 | 14.3 | 20.9 | 15.3 | 21.4 |
| mT5-base | 11.4 | 18.2 | 12.1 | 17.6 | 14.9 | 21.8 | 17.2 | 24.2 |
| mT5-large | 12.0 | 19.1 | 12.7 | 18.3 | 15.4 | 22.6 | 16.9 | 23.7 |
| mBART50 | 17.8 | **25.2** | 20.4 | 25.6 | 23.7 | 29.2 | 28.2 | 33.1 |
| mBART50-8k | **18.8** | 24.9 | **23.9** | **29.2** | **26.4** | **31.8** | **29.3** | **33.9** |

Table 13: **Present** keyphrases prediction results for the multilingual setting. Languages with a star ($*$) are unsupported by mBART50. Croatian with a dagger (†) is unsupported by mT5.

| Model | Bulgarian * | | Croatian † | | Czech | | Danish * | | Dutch | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mT5-small | 0.9 | 1.6 | 1.2 | 1.7 | 6.0 | 9.5 | 0.8 | 1.3 | 0.9 | 1.4 |
| mT5-base | 1.6 | 2.8 | 2.4 | 3.7 | 6.1 | 9.6 | 1.4 | 2.2 | 1.7 | 2.8 |
| mT5-large | 1.7 | 2.9 | 1.7 | 2.6 | 6.5 | 10.3 | 1.2 | 2.0 | 1.8 | 3.0 |
| mBART50 | **2.6** | 3.4 | **3.6** | **4.1** | 11.0 | 13.7 | **2.5** | **3.2** | 2.8 | 3.5 |
| mBART50-8k | 2.5 | **3.5** | 3.2 | 3.8 | **11.7** | **14.6** | 1.9 | 2.6 | **2.9** | **3.7** |

| Model | English | | Estonian | | Finnish | | French | | German | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mT5-small | 1.0 | 1.8 | 1.0 | 2.0 | 4.3 | 6.5 | 2.6 | 4.5 | 1.2 | 2.0 |
| mT5-base | 2.5 | 4.2 | 1.8 | 3.5 | 5.6 | 8.2 | 3.0 | 5.1 | 2.3 | 3.9 |
| mT5-large | 2.3 | 4.0 | 1.9 | 3.6 | 5.4 | 7.9 | 3.0 | 5.0 | 2.3 | 3.9 |
| mBART50 | 3.7 | 5.0 | **3.0** | **4.7** | 10.2 | 11.6 | 7.0 | 8.7 | **3.6** | 4.9 |
| mBART50-8k | **3.8** | **5.1** | **3.0** | 4.5 | **11.4** | **13.1** | **7.7** | **9.6** | **3.6** | **5.0** |

| Model | Greek * | | Hungarian * | | Irish * | | Italian | | Latvian | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mT5-small | 6.2 | 9.0 | 6.0 | 9.2 | 0.0 | 0.0 | 0.9 | 1.6 | 0.6 | 1.1 |
| mT5-base | 7.0 | 10.2 | 6.0 | 9.5 | 0.0 | 0.0 | 1.6 | 2.8 | 1.6 | 2.6 |
| mT5-large | 6.9 | 10.0 | 6.5 | 10.3 | 0.0 | 0.0 | 1.6 | 2.7 | 1.5 | 2.5 |
| mBART50 | 8.7 | 10.6 | 8.0 | 10.4 | 0.0 | 0.0 | **3.0** | **3.7** | 2.7 | 3.5 |
| mBART50-8k | **9.2** | **11.1** | **8.9** | **11.3** | 0.0 | 0.0 | 2.8 | 3.6 | **2.8** | **3.6** |

| Model | Lithuanian | | Maltese * | | Polish | | Portuguese | | Romanian | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mT5-small | 1.1 | 1.9 | 3.2 | 4.9 | 3.6 | 5.4 | 2.5 | 4.1 | 1.4 | 2.4 |
| mT5-base | 2.0 | 3.4 | 3.9 | 6.1 | 5.0 | 7.5 | 3.5 | 5.7 | 2.4 | 4.0 |
| mT5-large | 2.1 | 3.5 | 4.0 | 6.3 | 4.9 | 7.2 | 3.3 | 5.4 | 2.5 | 4.1 |
| mBART50 | 3.1 | 4.2 | **6.0** | **7.7** | 9.9 | 11.7 | 4.4 | 5.5 | 4.2 | 5.4 |
| mBART50-8k | **3.5** | **4.7** | 5.4 | 7.1 | **11.2** | **13.2** | **5.1** | **6.5** | **4.4** | **5.7** |

| Model | Slovak * | | Slovene | | Spanish | | Swedish | |
|---|---|---|---|---|---|---|---|---|
| | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* | *F1@5* | *F1@M* |
| YAKE | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mT5-small | 6.2 | 9.8 | 4.9 | 8.3 | 1.4 | 2.4 | 1.1 | 1.9 |
| mT5-base | 6.8 | 10.7 | 5.6 | 9.4 | 2.1 | 3.5 | 2.3 | 3.9 |
| mT5-large | 7.0 | 11.0 | 6.1 | 10.3 | 2.2 | 3.6 | 2.0 | 3.3 |
| mBART50 | 12.8 | 15.7 | 8.4 | 11.1 | 3.7 | **4.8** | 3.3 | 4.2 |
| mBART50-8k | **13.0** | **16.0** | **8.6** | **11.8** | **3.6** | **4.8** | **3.7** | **5.0** |

Table 14: **Absent** keyphrases prediction results for the multilingual setting. Languages with a star (∗) are unsupported by mBART50. Croatian with a dagger (†) is unsupported by mT5.

## H Random Split

As discussed previously, we perform a temporal split on the data to better reflect the reality of temporal concept shifting as well as new languages being introduced in the EU. However, a random split might be better in some cases, such as if we train a model for short-time use, and are willing to recollect data regularly to insure performance remains as predicted. In short, a random split can only procure guarantee about its performance as long as the real world data remains similar to the training data, while a temporal split give a view of the performance while accounting for changes in the real world data.

Still, to better allow practitioners to use data as they desire, we released a random split (https://huggingface.co/datasets/NCube/europa-random-split). The training, validation and test sets have respectively 10 003, 3 391 and 4 439 judgments. These figures are the same as in the chronologically split dataset. However, the number of instances per set in the random split setting differs since each set has the same distribution in terms of languages. Consequently, each set contains 159 306, 53 943 and 71 708 instances respectively. The results achieved by mBART50 are shown in Table 15. As expected, results are overall much higher since all training, validation and testing sets have the same distribution in terms of languages and vocabulary. We want to emphasize once again however that due to the fast changing nature of legal NLP, these performances are only valid for a short time after the data collection.

| Language | F1 Present | | | F1 Absent | | | MAP |
|---|---|---|---|---|---|---|---|
| | *@5* | *@10* | *@M* | *@5* | *@10* | *@M* | *@50* |
| Weighted Avg. | 30.4 | 20.0 | 41.5 | 15.3 | 10.3 | 18.1 | 20.6 |
| Unweighted Avg. | 30.6 | 20.2 | 41.3 | 15.2 | 10.3 | 17.9 | 20.7 |
| French | 30.2 | 19.7 | 41.2 | 14.9 | 9.8 | 17.8 | 20.2 |
| German | 30.0 | 19.8 | 40.3 | 14.2 | 9.3 | 18.4 | 19.8 |
| English | 32.8 | 22.0 | 43.6 | 9.9 | 6.5 | 13.3 | 19.0 |
| Italian | 27.7 | 18.2 | 38.7 | 16.3 | 10.7 | 20.0 | 19.4 |
| Dutch | 31.1 | 20.2 | 43.0 | 12.9 | 8.5 | 15.9 | 18.9 |
| Greek | 12.6 | 7.8 | 21.7 | 15.9 | 11.1 | 17.9 | 12.0 |
| Danish | 27.8 | 18.3 | 37.5 | 13.9 | 9.3 | 18.0 | 18.2 |
| Portuguese | 27.4 | 17.7 | 39.5 | 14.5 | 9.6 | 16.9 | 17.8 |
| Spanish | 30.1 | 19.7 | 41.2 | 14.1 | 9.2 | 17.2 | 19.2 |
| Swedish | 34.8 | 23.4 | 44.7 | 10.5 | 7.1 | 13.7 | 21.1 |
| Finnish | 28.2 | 17.7 | 42.4 | 17.9 | 12.3 | 21.0 | 20.1 |
| Lithuanian | 38.5 | 26.0 | 49.0 | 12.1 | 8.2 | 14.8 | 24.0 |
| Estonian | 40.2 | 28.0 | 48.1 | 9.8 | 6.6 | 12.4 | 24.8 |
| Czech | 28.9 | 18.6 | 41.8 | 24.4 | 16.9 | 26.6 | 24.5 |
| Hungarian | 27.1 | 17.2 | 40.3 | 18.4 | 12.5 | 21.6 | 20.1 |
| Latvian | 38.7 | 26.2 | 48.8 | 11.4 | 7.7 | 14.1 | 23.4 |
| Slovenian | 33.4 | 22.6 | 43.3 | 17.6 | 11.9 | 21.1 | 24.0 |
| Polish | 27.4 | 17.2 | 41.0 | 24.3 | 17.4 | 25.6 | 24.3 |
| Maltese | 28.1 | 17.9 | 40.5 | 18.3 | 12.6 | 20.5 | 19.8 |
| Slovak | 31.3 | 20.0 | 45.0 | 22.0 | 15.2 | 24.0 | 24.2 |
| Romanian | 35.2 | 23.5 | 46.5 | 19.4 | 13.2 | 21.5 | 25.6 |
| Bulgarian | 35.4 | 23.8 | 46.3 | 11.2 | 7.6 | 13.1 | 21.9 |
| Croatian | 40.8 | 28.6 | 46.9 | 13.7 | 9.3 | 13.9 | 26.4 |
| Irish | 15.9 | 11.5 | 20.5 | 7.7 | 5.1 | 10.2 | 8.9 |

Table 15: Results on a random split of the data, per language, achieved by mBART50.