Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations

Ilias Chalkidis [† ‡] **Manos Fergadiotis** [† ‡] **Nikolaos Manginas** [†]

Eva Katakalou [♭ *] **Prodromos Malakasiotis** [† ‡]

[†] EY AI Centre of Excellence in Document Intelligence, NCSR "Demokritos"
[‡] Department of Informatics, Athens University of Economics and Business
[♭] Department of International, European and Area Studies, Panteion University

[Ilias.Chalkidis, Fergadiotis.Manos]@ey.com
[Nikolaos.Manginas, Prodromos.Malakasiotis]@ey.com
[ichalkidis, mfergadiotis]@iit.demokritos.gr
[nmanginas, pmalakasiotis]@iit.demokritos.gr
e.katakalou@panteion.gr

Abstract

Major scandals in corporate history have urged the need for *regulatory compliance*, where organizations need to ensure that their controls (processes) comply with relevant laws, regulations, and policies. However, keeping track of the constantly changing legislation is difficult, thus organizations are increasingly adopting Regulatory Technology (RegTech) to facilitate the process. To this end, we introduce *regulatory information retrieval* (REG-IR), an application of *document-to-document information retrieval* (DOC2DOC IR), where the query is an entire document making the task more challenging than traditional IR where the queries are short. Furthermore, we compile and release two datasets based on the relationships between EU directives and UK legislation. We experiment on these datasets using a typical two-step pipeline approach comprising a pre-fetcher and a neural re-ranker. Experimenting with various pre-fetchers from BM_{25} to k nearest neighbors over representations from several BERT models, we show that fine-tuning a BERT model on an in-domain classification task produces the best representations for IR. We also show that neural re-rankers underperform due to *contradicting* supervision, i.e., similar query-document pairs with opposite labels. Thus, they are biased towards the pre-fetcher's score. Interestingly, applying a date filter further improves the performance, showcasing the importance of the time dimension.

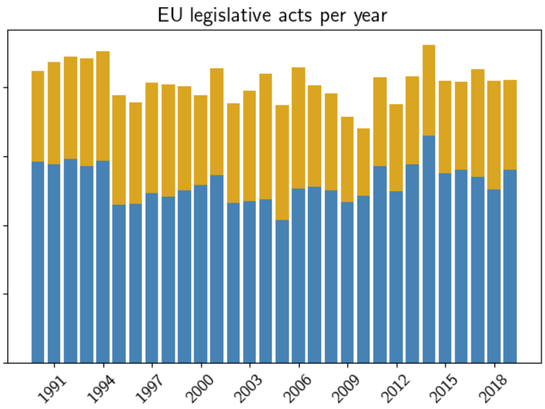

Figure 1: Number of legislative acts issued by the EU per year. The gold color of the bars indicates how many of the published acts are amendments to older ones.

1 Introduction

Major scandals in corporate history, from Enron to Tyco International, Olympus, and Tesco,[1] have led to the emergence of stricter regulatory mandates and highlighted the need for *regulatory compliance* where organizations need to ensure that they comply with relevant laws, regulations, and policies (Lin, 2016). However, keeping track of the constantly changing legislation (Figure 1) is hard, thus organizations are increasingly adopting Regulatory Technology (RegTech) to facilitate the process.

Typically, a compliance regimen includes three distinct but related types of measures, *corrective*, *detective*, and *preventive* (Sadiq and Governatori,

[*] The contribution of Ms. Eva Katakalou was restricted to the creation and the validation of the datasets as well as to the authoring of the corresponding parts of the manuscript.

[1] www.theguardian.com/business/2015/jul/21/the-worlds-biggest-accounting-scandals-toshiba-enron-olympus

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3498–3511
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

2015). Corrective measures are usually undertaken when new regulations are introduced to update existing controls. Detective measures, ensure "after-the-fact" compliance, i.e., following a procedure, a manual or automated check is carried out, to ensure that every step of the procedure complied with the corresponding regulations. Finally, preventive measures ensure compliance "by design", i.e., during the creation of new controls. All types of measures include an underlying information retrieval (IR) task, where laws need to be retrieved given a control or vice versa. We identify two use cases:

1. *Given a new law retrieve all the controls of the organization affected by this law*. The organization can then apply corrective measures to ensure compliance for these controls.
2. *Given a control retrieve all relevant laws the control should comply with*. This is useful for ensuring compliance after a procedure has been carried out (detective measures) or when creating new controls (preventive measures).

Regulatory information retrieval (REG-IR), similarly to other applications of *document-to-document* (DOC2DOC) IR, is much more challenging than traditional IR where the query typically contains a few informative words and the documents are relatively small (Table 1). In DOC2DOC IR the query is a long document (e.g., a regulation) containing thousands of words, most of which are uninformative. Consequently, matching the query with other long documents where the informative words are also sparse, becomes extremely difficult.

Although legislation is available, organizations' controls are strictly private and very hard to obtain. Fortunately, the European Union (EU) has a legislation scheme analogous to regulatory compliance for organizations. According to the Treaty on the Functioning of the European Union (TFEU),[2] all published EU *directives* must take effect at the national level. Thus, all EU member states must adopt a law to transpose a newly issued directive within the period set by the directive (typically 2 years). Notably, the United Kingdom (UK) having a high compliance level with the EU (Figure 2),[3] is a good test-bed for REG-IR. Thus we compile and release two datasets for REG-IR, EU2UK and UK2EU, containing EU directives and UK regulations, which can serve both as queries and documents under the ground truth assumption that a UK law is relevant to the EU directives it transposes and vice versa.

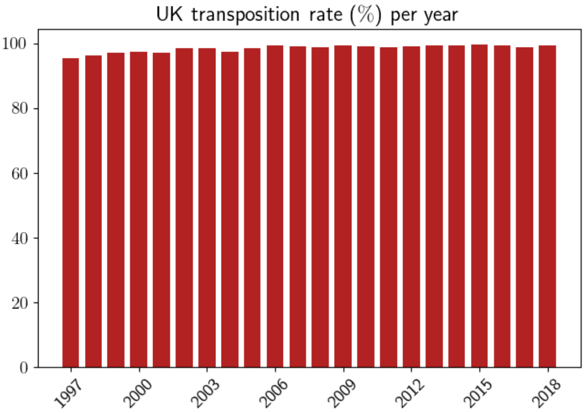

Figure 2: The percentage of EU directives transposed by UK legislation per year. Over 98% of the published EU directives have been transposed.

Dataset	Domain	$\tilde{q}$	$\tilde{d}$
IR datasets in the literature			
TREC ROBUST (Voorhees, 2005)	News	3 / 14	254
BIOASQ (Tsatsaronis et al., 2015)	Biomedical	9	197
IR datasets with verbose queries			
GOV2 (Clarke et al., 2004)	Web	11 / 57	682
WT10G (Chiang et al., 2005)	Web	11 / 35	457
Regulatory Compliance datasets			
EU2UK (ours)	Law	2,642	1,849
UK2EU (ours)	Law	1,849	2,642

Table 1: Statistics for query and document length for IR datasets used in literature.

Since REG-IR is a new task, our starting point is the two-step pipeline approach followed by most modern neural information retrieval systems (Guo et al., 2016; Hui et al., 2017; McDonald et al., 2018). First, a conventional IR system (*pre-fetcher*) retrieves the k most prominent documents. Then a neural model attempts to rank relevant documents higher than irrelevant ones. In most approaches, the pre-fetcher is based on Okapi BM25 (Robertson et al., 1995), a bag-of-words scoring function that does not consider possible synonyms or contextual information. To overcome the first limitation, we follow Brokos et al. (2016) who employed k nearest neighbors over tf-idf weighted centroids of word embeddings, without however improving the results, probably because the centroids are noisy considering many uninformative words. Furthermore, we employ BERT (Devlin et al., 2019) to extract contextualized representations for queries and documents but again the results are worse than BM25. We also experiment with S-BERT (Reimers

[2] Articles 291 (1) and 288 paragraph 3.

[3] Data for Figures 1 and 2 obtained from ec.europa.eu/internal_market/scoreboard/performance_by_governance_tool/eu_pilot.

Query: DIRECTIVE 2006/66/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 September 2006 on batteries and accumulators and waste batteries and accumulators and repealing Directive 91/157/EEC

BM$_{25}$ rank	Relevant	Document title
1	No	The Batteries and Accumulators (Placing on the Market) (Amendment) Regulations 2012
2	No	The Batteries and Accumulators (Containing Dangerous Substances) (Amendment) Regulations 2000
3	No	The Batteries and Accumulators (Placing on the Market) (Amendment) Regulations 2015
4	No	The Batteries and Accumulators (Containing Dangerous Substances) Regulations 1994
5	No	The Waste Batteries and Accumulators (Amendment) Regulations 2015
6	Yes	The Waste Batteries and Accumulators Regulations 2009
12	Yes	The Batteries and Accumulators (Placing on the Market) Regulations 2008

Table 2: Example from the EU2UK dataset where the retrieved UK laws are ranked by BM$_{25}$. The top-5 documents seem similar to the query but are not relevant. Documents ranked 1st, 3rd, and 5th are amendments of the relevant documents, i.e., UK laws that transpose the query.

and Gurevych, 2019) and LEGAL-BERT (Chalkidis et al., 2020), a model specialized in the legal domain. Both models perform better than BERT but are still worse than or comparable to BM$_{25}$. The inability of BERT-based models motivated us to find an auxiliary task that will result in better representations for REG-IR. Following Chalkidis et al. (2019), we fine-tune BERT to predict EUROVOC concepts that describe the core subjects of each text. As expected this model (C-BERT) is the best pre-fetcher by a large margin in EU2UK, while being comparable to BM$_{25}$ in UK2EU. To summarize, our contributions are:

(a) We introduce REG-IR, an application of DOC2DOC IR, which is a new family of IR tasks, where both queries and documents are long typically containing thousands of words.

(b) We compile and release the two first publicly available datasets, EU2UK and UK2EU, suitable for REG-IR and DOC2DOC IR in general.[4]

(c) We show that fine-tuning BERT on an in-domain classification task produces the best document representations with respect to IR and improves pre-fetching results.

2 Datasets curation

2.1 Data sources

EU/UK Legislation: We have downloaded approx. 56K pieces of EU legislation (approx. 3.9K directives), from the EURLEX portal.[5] EU laws are 2,642 words long on average and are structured in three major parts: the *title* (Table 2, query), the *recitals* consisting of references in the legal background of the act, and the *main body*. We have also downloaded approx. 52K UK laws, publicly available from the official UK legislation portal.[6] UK laws are 1,849 words long on average and contain the *title* (Table 2, document title) and the *main body*.

Transpositions: We have retrieved all transposition relations (approx. 3.7K) between EU directives and UK laws from the CELLAR database. CELLAR only provides the mapping between the CELLAR ids of EU directives and the title of each UK law. Therefore we aligned the CELLAR ids with the official UK ids based on the law title.[7] One or more UK laws may transpose one or more EU directives.

2.2 Datasets compilation

Let $\mathcal{E}, \mathcal{U}$ be the sets of EU directives and UK laws, respectively. We define REG-IR as the task where the query q is a document, e.g, an EU directive, and the objective is to retrieve a set of relevant documents, $\mathcal{R}_q$, from the pool of all available documents, e.g., all UK laws. We create two datasets:

EU2UK: $q \in \mathcal{E}, \mathcal{R}_q = \{r_i \colon r_i \in \mathcal{U}, r_i \xrightarrow{\text{transposes}} q\}$.

UK2EU: $q \in \mathcal{U}, \mathcal{R}_q = \{r_i \colon r_i \in \mathcal{E}, q \xrightarrow{\text{transposes}} r_i\}$.

Table 3 shows the statistics for the two datasets, which are split in three parts, *train*, *development*, and *test*, retaining a chronological order for the queries. EU2UK has a much larger pool of available documents than UK2EU (52.5K vs. 3.9K) which may impose an extra difficulty during retrieval. More importantly, the average number of relevant documents per query is small (at most 2) for both datasets, as our ground truth assumption is strict, i.e., relevant documents are those linked to the query with a transposition relation. Also, EU legislation is frequently amended (Figure 1) which also

[4] The datasets are available at `https://archive.org/details/eacl2021_regir_datasets`.

[5] `eur-lex.europa.eu`

[6] `legislation.gov.uk`

[7] See Appendix A for details on the dataset curation.

Dataset	Documents in pool	Train		Development		Test	
		Queries	Avg. relevant	Queries	Avg. relevant	Queries	Avg. relevant
EU2UK	52,515	1,400	1.79	300	2.09	300	1.74
UK2EU	3,930	1,500	1.90	300	1.46	300	1.29

Table 3: Detailed statistics for EU2UK and UK2EU. Both datasets have relatively small number of relevant documents while EU2UK has also large pool which may impose extra difficulties in the retrieval.

imposes difficulty in the retrieval task. Let $d_1 \in \mathcal{E}$ be a directive transposed by $u_1 \in \mathcal{U}$ and $d_2 \in \mathcal{E}$ be a directive amending d_1. The UK must adopt a law, u_2, to transpose d_2. Both d_2 and u_2 cover similar concepts to those of d_1 (d_2 is an amendment and u_2 must comply with d_2), but, strictly speaking u_2 is relevant only to d_2. Table 2 shows an example from EU2UK, where the top-5 documents seem very similar to the query but are not considered relevant. Note that the documents ranked 1st, 3rd and 5th, are amendments of the relevant documents.

3 IR pipelines

Modern neural IR systems usually follow a two-step pipeline approach. First, a conventional IR system (*pre-fetcher*) retrieves the top-k most prominent documents aiming to maximize its recall. Then a neural model attempts to re-rank the documents by scoring relevant higher than irrelevant ones. While this configuration is widely adopted in literature, the re-ranking step could be omitted provided an effective pre-fetching mechanism, i.e., the pre-fetcher will act as an end-to-end IR system.

3.1 Document pre-fetching

Okapi BM25 (Robertson et al., 1995) is a bag-of-words scoring function estimating the relevance of a document d to a query q, based on the query terms appearing in d, regardless their proximity within d:

$$\sum_{i=1}^{n} \mathrm{idf}(q_i) \cdot \frac{\mathrm{tf}(q_i, d) \cdot (k_1 + 1)}{\mathrm{tf}(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{L}{\bar{L}}\right)} \quad (1)$$

where q_i is the i-th query term, with $\mathrm{idf}(q_i)$ inverse document frequency and $\mathrm{tf}(q_i, d)$ term frequency. L is the length of d in words, $\bar{L}$ is the average length of the documents in the collection, k_1 is a parameter that favors high tf scores and b is a parameter penalizing long documents.[8]

W2V-CENT: Following Brokos et al. (2016), we represent query/document terms with pre-trained embeddings. For each query/document we calculate the tf-idf weighted centroid of its embeddings:

$$\mathrm{cent}(t) = \frac{\sum_{i=1}^{l} \boldsymbol{x}_i \cdot \mathrm{tf}(x_i, t) \cdot \mathrm{idf}(x_i)}{\sum_{i=1}^{l} \mathrm{tf}(x_i, t) \cdot \mathrm{idf}(x_i)} \quad (2)$$

where t is a text (query or document) and x_i is the i-th text term with embedding $\boldsymbol{x}_i$. The documents are ranked, with respect to the query, by a k nearest neighbours (kNN) algorithm with cosine distance:

$$\cos_{\mathrm{d}}(q, d) = 1 - \frac{\mathrm{cent}(q) \cdot \mathrm{cent}(d)}{\|\mathrm{cent}(q)\| \cdot \|\mathrm{cent}(d)\|} \quad (3)$$

BERT, similarly to W2V-CENT, relies in pre-trained representations which now are extracted from BERT, thus being context-aware. A text can be represented by its `[cls]` token or by the centroid of its token embeddings. In the latter case the embeddings can be extracted from any of the 12 layers of BERT.[9] Note that the texts in our datasets do not entirely fit in BERT. We thus split them into c chunks (2 to 3 per text) and pass each chunk through BERT to obtain a list of token embeddings per layer (i.e, the concatenation of c token embeddings lists) or c `[cls]` tokens. The final representation is either the centroid of the token embeddings or the centroid of the `[cls]` tokens.

S-BERT (Reimers and Gurevych, 2019) is a BERT model fine-tuned for NLI. According to the authors, training S-BERT for NLI results in better representations than BERT for tasks involving text comparison, like IR. We use the same setting as in BERT.

LEGAL-BERT: Our datasets come from the legal domain which has distinct characteristics compared to generic corpora, such as specialized vocabulary, particularly formal syntax, semantics based on extensive domain-specific knowledge, etc., to the extent that legal language is often classified as a 'sublanguage' (Tiersma, 1999; Williams, 2007; Haigh, 2018). BERT and S-BERT were trained on generic corpora and may fail to capture the nuances of legal language. Thus we used a BERT model further pre-trained on EU legislation (Chalkidis et al., 2020), dubbed here LEGAL-BERT, in a similar fashion.

[8] We use *elastic*, a widely used IR engine with the BM25 scoring function. See www.elastic.co/.

[9] BERT is not fine-tuned during this process.

C-BERT: EU laws are annotated with EUROVOC concepts covering the core subjects of EU legislation (e.g., environment, trade, etc.). Our intuition is that a UK law transposing an EU directive will most probably cover the same subjects. Thus we expect that a BERT model, fine-tuned to predict EUROVOC concepts, will learn rich representations describing these concepts which may be useful for pre-fetching. We fine-tune BERT following Chalkidis et al. (2019)[10] and use the resulting model to extract query and document representations similarly to the previous BERT-based methods.

ENSEMBLE is simply a combination of our best two pre-fetchers, C-BERT and BM$_{25}$:

$$\text{ENS}(q,d) = \alpha \cdot \text{CB}(q,d) + (1-\alpha) \cdot \text{BM}_{25}(q,d) \quad (4)$$

where CB is the score of C-BERT and α is tuned on development data and the scores of the pre-fetchers are normalized in $[0,1]$.

3.2 Document re-ranking

Modern neural re-rankers operate on pairs of the form (q,d) to produce a relevance score, $\text{rel}(q,d)$, for a document d with respect to a query q. Note, however, that the main objective is to rank relevant documents higher than irrelevant. Thus, during training the loss is calculated as:

$$\mathcal{L} = \max(0, 1 - \text{rel}(q, d^{+}) + \text{rel}(q, d^{-})) \quad (5)$$

where d^{+} is a relevant document and d^{-} is an irrelevant document. We have experimented with several neural re-ranking methods each having a function that produces a relevance score s_r for each of the top-k documents returned by the best pre-fetcher. The final relevance score of a document is calculated as: $\text{rel}(q,d) = w_r \cdot s_r + w_p \cdot s_p$, where s_p is the normalized score of the pre-fetcher and w_s, w_p are learned during training.

Given the concerns on the strictness of the ground truth assumption raised in Section 2.2, we hypothesize that re-rankers will eventually over-utilize the pre-fetcher score, s_p, when calculating document relevance, $\text{rel}(q,d)$. As shown in Table 2, in many cases both relevant and irrelevant documents may have high similarity with the query. This in turn may confuse and therefore degenerate the re-ranker's term matching mechanism, i.e., MLPs or CNNs over term similarity matrices.

DRMM (Guo et al., 2016) uses pre-trained word embeddings to represent query and document terms. A histogram captures the cosine similarities of a query term, q_i, with all the terms of a particular document. Then an MLP consumes the histograms to produce a document-aware score for each q_i, which is weighted by a gating mechanism assessing the importance of q_i. The sum of the weighted scores is the relevance score of the document. A caveat of DRMM is that it completely ignores the context of the terms which could be of particular importance in our datasets where texts are long.

PACRR (Hui et al., 2017) represents query and document terms with pre-trained embeddings and calculates a matrix S containing the cosine similarities of all query-document term pairs. A row-wise k-max pooling operation on S keeps the highest similarities per query term (matrix S_k). Then, wide convolutions of different kernel (filter) sizes ($n \times n$) with multiple filters per size are applied on S. Each filter of size $n \times n$ attempts to capture n-gram similarities between queries and documents. A max-pooling operation keeps the strongest signals across filters and a row-wise k-max pooling keeps the strongest signals per query n-gram, resulting in the matrix $S_{n,k}$. Subsequently, a row-wise concatenation of S_k with all $S_{n,k}$ matrices (for different values of n) is performed and a column containing the softmax-normalized idf scores of the query terms is concatenated to the resulting matrix (S_{sim}). In effect, each row of the matrix contains different n-gram based similarity views of the corresponding query term, q_i, along with an idf-based importance score. The relevance score is produced as the last hidden state of an LSTM with one hidden unit, which consumes the rows of S_{sim}. PACRR tries to take into account the context of the query and document terms using n-grams but this context sensitivity is weak and we do not expect much benefits in our datasets which contain long texts.

BERT-based re-rankers: Recent work tries to exploit BERT to improve re-ranking. Following MacAvaney et al. (2019), we use DRMM and PACRR on top of contextualized BERT embeddings derived from BERT. Based on the results of Figure 4, we use C-BERT as the most promising BERT model. We call these two models C-BERT-DRMM and C-BERT-PACRR. We also experiment with two settings depending on whether C-BERT weights are updated (*tuned*) or not (*frozen*) during training.

[10] We use all EU laws excluding EU directives that exist in our development and test sets.

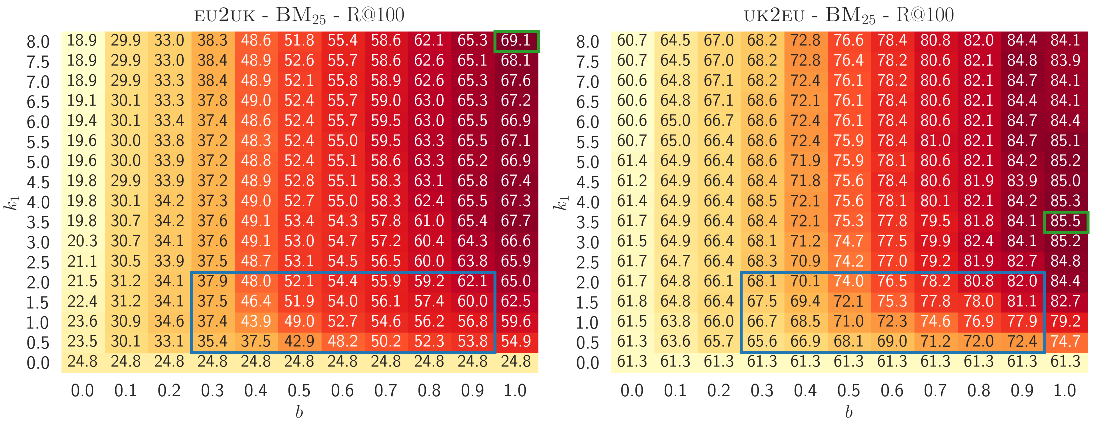

Figure 3: Heatmaps showing R@100 for different values of k_1 and b on EU2UK (left) and UK2EU (right). The selected optimal values (green boxes) are outside the proposed ranges in the literature (blue boxes).

4 Experimental setup

4.1 Pre-trained resources

As several methods rely on word embeddings, we trained a new WORD2VEC model (Mikolov et al., 2013) in both corpora (EU and UK legislation) to better accommodate legal language. Preliminary experiments showed that domain-specific embeddings perform better than generic 200-dimensional GloVe embeddings (Pennington et al., 2014) in development data (EU2UK: 66.5 vs. 59.3 at R@100 and UK2EU: 72.6 vs. 69.8 at R@100).[11]

All BERT (pre-fetching) encoders and BERT-based re-rankers use the -BASE version, i.e., 12 layers, 768 hidden units and 12 attention heads, similar to the one of Devlin et al. (2019).[12]

4.2 Pre-processing - document denoising

One of the major challenges in DOC2DOC IR, as opposed to traditional IR, is the length of the queries and the documents which may induce noise (many uninformative words) during retrieval. Thus we applied several filters (stop-word, punctuation and digits elimination) on both queries and documents and reduced their length by approx. 55% (778 words for UK laws and 1,222 words for EU directives on average). Further on, we filtered both queries and documents by eliminating words with idf score less than the average idf score of the stop-words. Our intuition is that words (e.g., regulation, EU, law, etc.) with such a small idf score are uninformative. Still, the texts are much longer (387 words for UK laws and 631 words for EU directives on average) than the queries used in traditional IR (Table 1). As an alternative to drastically decrease the query size, we experimented with using only the title of a legislative act as a query but the results were worse, i.e., approx. 5-20% lower R@100 on average across datasets, indicating that the full-text is more informative, although the information is sparse. Hence, we only consider the full-text, including the title, for the rest of the experiments.

4.3 Evaluation measures

Pre-fetching aims to bring all the relevant documents in the top-k, thus we report R@k. We observe that for $\mathrm{k} > 100$ the best pre-fetchers have not significant gains in performance in development data, thus we select $\mathrm{k} = 100$, as a reasonable threshold.[13] For re-ranking we report R@20, nDCG@20 and R-Precision (RP) following the literature (Manning et al., 2009). We report the average and standard deviation across three runs considering the best set of hyper-parameters on development data for neural re-rankers.

4.4 Tuning BM$_{25}$: The case of DOC2DOC IR

The effectiveness of BM$_{25}$ is highly dependant on properly selecting the values of k_1 and b. In traditional (ad-hoc) IR, k_1 is typically evaluated in the range$[0, 3]$ (usually $k_1 \in [0.5, 2.0]$); b needs to be in $[0, 1]$ (usually $b \in [0.3, 0.9]$) (Taylor et al., 2006; Trotman et al., 2014; Lipani et al., 2015). As a general rule of thumb BM$_{25}$ with k_1=1.2 and b=0.75 seems to give good results in most cases (Trotman et al., 2014). We observe that in the case of DOC2DOC IR where the queries are much longer, the optimal values are outside the proposed ranges

[11] See also the discussion for legal language in Section 3.1.

[12] See Appendix B for more details.

[13] See Appendix A.3 for an extended ($\mathrm{k} \in [0, 2000]$) performance evaluation on pre-fetching.

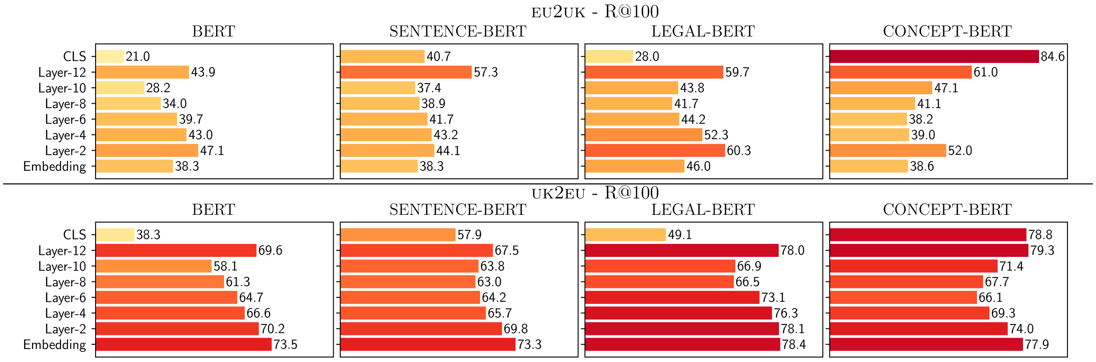

Figure 4: Heatbars showing R@100 (on development data) for text representations extracted from different layers of the various BERT-based pre-fetchers we experimented with.

(Figure 3). In both datasets the optimal values for k_1 and b are relatively high, favoring terms with high tf, while penalizing long documents. In effect BM$_{25}$ uses k_1 and b as a denoising regularizer to over-utilize highly frequent query terms normalized by document length.

4.5 Extracting representations from BERT

Recently there has been a lot of research on understanding the effectiveness of BERT's different layers (Liu et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Goldberg, 2019; Kovaleva et al., 2019; Lin et al., 2019). Figure 4 shows heatbars comparing representations extracted from different layers of the various BERT-based pre-fetchers we experimented with.[14] LEGAL-BERT and C-BERT which have been adapted in the legal domain perform much better than BERT and S-BERT which were trained on generic corpora. An interesting observation is that the [cls] token is a powerful representation only in C-BERT where it was trained to predict EUROVOC concepts. Also, in UK2EU the embedding layer produces the best representations in all BERT variants except C-BERT, where the embedding layer achieves comparable results to the top-2 representations ([cls], Layer-12). This is an indication that the context in this dataset is not as important as in EU2UK.

4.6 Implementation details

All neural models were implemented using the Tensorflow 2 framework. Hyper-parameters were tuned on development data, using early stopping and the Adam optimizer (Kingma and Ba, 2015).

Method	EU2UK	UK2EU
	R@100	R@100
BM$_{25}$ (Robertson et al., 1995)	57.5	<u>93.7</u>
W2V-CENT (Brokos et al., 2016)	50.6	88.2
BERT (Devlin et al., 2019)	54.0	85.1
S-BERT (Reimers and Gurevych, 2019)	57.7	84.8
LEGAL-BERT (Chalkidis et al., 2020)	57.6	90.1
C-BERT (ours)	<u>83.8</u>	<u>92.9</u>
ENSEMBLE (BM$_{25}$ + C-BERT)	**86.5**	**95.0**

Table 4: Pre-fetching results across test datasets.

5 Experimental results

Pre-fetching: Table 4 shows R@100 on the test datasets for the various pre-fetchers considered. On EU2UK, C-BERT is the best method by a large margin, followed by S-BERT and LEGAL-BERT, verifying our assumption that the concept classification task is a good proxy for obtaining rich representations with respect to IR. Both S-BERT and LEGAL-BERT are better than BERT for different reasons. LEGAL-BERT was adapted to the legal domain and is, therefore, able to capture the nuances of the legal language. S-BERT was trained to produce representations suitable for comparing texts with cosine similarity, a task highly related to IR. Nonetheless, having been trained on generic corpora with small texts, it performs much worse than C-BERT. Interestingly, BM$_{25}$ is comparable to both S-BERT and LEGAL-BERT despite its simplicity. As expected, combining C-BERT with BM$_{25}$ further improves the results. In UK2EU R@100 is much higher compared to EU2UK probably because of the shortest queries. Also, as discussed in Section 4.5, the contextual information is not so critical in this dataset, thus we expect the context unaware BM$_{25}$ and W2V-CENT to perform well. Indeed, BM$_{25}$ achieves the best results followed closely by C-BERT and LEGAL-BERT, while W2V-

[14] Recall that a text can be represented by its [cls] token or by the centroid of its token embeddings which can be extracted from any of the 12 layers of BERT.

Method	EU2UK					UK2EU				
	w_p	w_s	R@20	nDCG@20	RP	w_p	w_s	R@20	nDCG@20	RP
BM$_{25}$	-	-	45.8	34.4	25.5	-	-	87.5	66.8	**49.4**
C-BERT (ours)	-	-	55.7	37.9	21.8	-	-	79.7	53.0	33.1
ENSEMBLE (BM$_{25}$ + C-BERT)	-	-	54.1	43.1	29.6	-	-	88.0	**67.7**	49.3
+ DRMM	+1.1	-0.8	**59.9** (± 3.2)	41.7 (± 2.4)	24.3 (± 2.9)	+1.3	-0.8	86.3 (± 1.1)	61.6 (± 1.1)	40.1 (± 1.5)
+ PACRR	+4.2	+0.6	54.3 (± 0.2)	**43.3** (± 0.2)	**30.1** (± 0.4)	+4.0	+0.1	88.0 (± 0.0)	**67.7** (± 0.0)	49.3 (± 0.0)
+ C-BERT-DRMM *(frozen)*	+3.3	-1.6	57.9 (± 3.4)	43.1 (± 0.3)	27.3 (± 2.2)	+3.5	-1.0	88.3 (± 0.4)	67.3 (± 0.6)	48.5 (± 1.3)
+ C-BERT-PACRR *(frozen)*	+4.6	+0.9	54.1 (± 0.0)	43.1 (± 0.0)	29.6 (± 0.0)	+2.9	-0.9	**89.6** (± 0.4)	66.5 (± 0.5)	46.0 (± 0.9)
+ C-BERT-DRMM *(tuned)*	+1.9	-0.5	54.1 (± 0.0)	43.1 (± 0.0)	29.6 (± 0.0)	+1.2	+0.5	88.0 (± 0.0)	**67.7** (± 0.0)	49.3 (± 0.0)
+ C-BERT-PACRR *(tuned)*	+1.8	-0.6	54.1 (± 0.0)	43.1 (± 0.0)	29.6 (± 0.0)	+2.0	+2.1	88.0 (± 0.0)	**67.7** (± 0.0)	49.3 (± 0.0)
+ ORACLE	-	-	86.5	87.7	86.5	-	-	95.0	95.3	95.0
Applying date filtering on top of predictions										
Year range	±5 years					±15 years				
ENSEMBLE (BM$_{25}$ + C-BERT)	-	-	76.6	54.6	37.1	-	-	**86.2**	**68.2**	**50.0**
+ DRMM (*pre-filtering*)	+1.1	-0.8	**81.4**	**56.5**	35.4	+1.3	-0.8	85.3	62.6	42.3
+ DRMM (*post-filtering*)	+1.1	-0.8	75.7	49.2	31.1	+1.3	-0.8	83.6	63.5	44.2
+ PACRR (*pre-filtering*)	+4.2	+0.6	76.6	54.8	**37.6**	+4.0	+0.1	**86.2**	**68.2**	**50.0**
+ PACRR (*post-filtering*)	+4.2	+0.6	74.2	52.9	36.5	+4.0	+0.1	85.5	67.6	49.6

Table 5: Re-ranking results across test datasets. The upper zone shows the results of neural re-rankers on top of the best pre-fetchers with respect to (w_s, w_p). It also reports re-ranking results of the best pre-fetchers. The lower zone reports the re-ranking results after applying temporal filtering.

CENT outperforms S-BERT and BERT. Again the ENSEMBLE improves the results.

Re-ranking: Table 5 shows the ranking results on test data for EU2UK and UK2EU. We also report results for BM$_{25}$, C-BERT, ENSEMBLE and an ORACLE, which re-ranks the top-k documents returned by the pre-fetcher placing all relevant documents at the top. On EU2UK ENSEMBLE performs better than the other two pre-fetchers. Interestingly, neural re-rankers fall short on improving performance and are comparable (or even identical) with ENSEMBLE in most cases, possibly because very similar documents may be relevant or not (Section 2.2, Table 2), leading to *contradicting* supervision.[15] As we hypothesized (Section 3.2), re-rankers over-utilize the pre-fetcher score when calculating document relevance, as a defense mechanism (bias) against contradicting supervision, which eventually leads to the degeneration of the re-ranker's term matching mechanism. Inspecting the corresponding weights of the models, we observe that indeed $w_p >> w_s$ across all methods. This effect seems more intense in BERT-based re-rankers (C-BERT + DRMM or PACRR), especially those that fine-tune C-BERT, possibly because these models perform term matching considering sub-word units, instead of full words. In other words, relying on the neural relevance score (s_r) is catastrophic. Similar observations can be made for UK2EU. In both datasets all methods have a large performance gap compared to the ORACLE, indicating that there is still large room for improvement, possibly utilizing information beyond text.

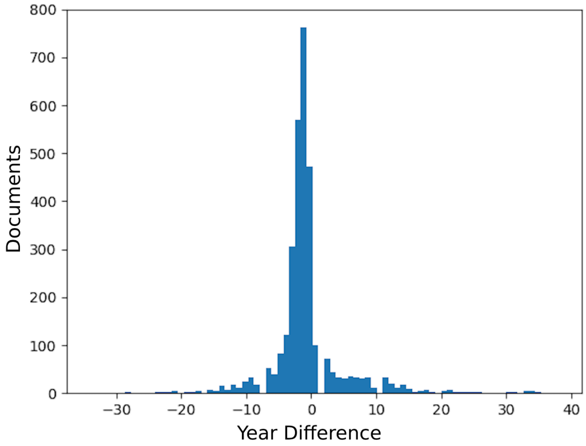

Figure 5: Relevant documents according to their chronological difference with the query on EU2UK development data.

Filtering by year: We have already highlighted the difficulties imposed to our datasets by the frequently amended EU directives (Section 2.2, Table 2). Also, recall that each EU directive defines a deadline (typically 2 years) for the transposition to take place. On the other hand, as we observe in Figure 5, EU directives may already be transposed by earlier legislative acts of member states (the member states act in a proactive manner), or they may delay the transposition for political reasons. In effect, the relevance of a document to a query depends both on the textual content and the time the laws were published. Thus, we filter out documents that are outside a predefined distance

[15] By *contradicting* supervision we mean similar training query-document pairs with opposite labels.

(in years) from the query in two ways, *pre-filtering* and *post-filtering*. Pre-filtering is applied to the pre-fetcher, i.e., prior to re-ranking, while post-filtering is applied after the re-ranking. Note that our main goal is to improve re-ranking. We thus apply the filtering scheme to the ENSEMBLE, DRMM and PACRR. The lower zone of Table 5 shows the results of the whole process. In EU2UK, the hardest out of the two datasets, the time filtering has a positive impact, improving the results by a large margin. On the other hand, filtering seems to have a minor effect in UK2EU.

5.1 EU2UK $\neq$ UK2EU

Across experiments, we observe that best practices vary between the EU2UK and UK2EU datasets. EU2UK benefits from C-BERT representations, while in UK2EU context-unaware and domain-agnostic BM$_{25}$ has comparable or better performance than C-BERT. Similarly, we observe that time filtering further improves the performance in EU2UK, while we have a contradicting effect in UK2EU. Given the overall results, we conclude the two datasets have quite different characteristics. Thus, it is important to consider both EU2UK and UK2EU independently, although one may initially consider them to be symmetric.

6 Related work

IR in the legal domain is widely connected with the Competition on Legal Information Extraction/Entailment (COLIEE). From 2015 to 2017 (Kim et al., 2015, 2016; Kano et al., 2017), the task was to retrieve Japanese Civil Code articles given a question, while in COLIEE 2018 and 2019 (Kano et al., 2018; Rabelo et al., 2019), the task was to retrieve supporting cases given a short description of an unseen case. However, the texts of these competitions are small compared to our datasets. Also, most submitted systems do not consider recent advances in IR, i.e, neural ranking models (Guo et al., 2016; Hui et al., 2017; McDonald et al., 2018; MacAvaney et al., 2019), which have recently managed to improve rankings of conventional IR, or end-to-end neural models which have recently been proposed (Fan et al., 2018; Khattab and Zaharia, 2020). Again, these end-to-end methods were applied on small texts. On the other hand, there has been some work trying to cope with larger queries, i.e., *verbose* or expanded queries, (Paik and Oard, 2014; Gupta and Bendersky, 2015; Cummins, 2016). Nonetheless, the considered queries are at most 60 tokens long, contrary to our datasets where, depending on the setting, the average query length is 1.8K or 2.6K tokens (Table 1). Neural methods greatly rely on text representations, thus Reimers and Gurevych (2019) proposed S-BERT which is trained to compare texts for an NLI task and could thus be used to extract representations suitable for IR. Towards the same direction, Chang et al. (2020) experimented with several auxiliary tasks to extract better representations. However, the latter two methods have been evaluated on datasets with much smaller texts than the ones we consider.

7 Conclusions and future work

We proposed DOC2DOC IR, a new family of IR tasks, where the query is an entire document, thus being more challenging than traditional IR. This family of tasks is particularly useful in regulatory compliance, where organizations need to ensure that their controls comply with the existing legislation. In the absence of publicly available DOC2DOC datasets, we compile and release two datasets, containing EU directives and UK laws transposing these directives. Experimenting with conventional (BM$_{25}$) and neural pre-fetchers we showed that a BERT model fine-tuned on an in-domain classification task, i.e., predict EUROVOC concepts, is by far the best pre-fetcher in our datasets. We also showed that neural re-rankers fail to improve the performance, as their term matching mechanisms degenerates, and over-utilize the pre-fetcher score. In the future, we would like to investigate alternatives in exploiting additional information that may be critical in the newly introduced tasks (EU2UK, UK2EU). In this direction naively utilizing chronological information leads to vast performance improvement in EU2UK dataset. One possible direction is to model the cross-document relations (e.g., amendments) using Graph Convolutional Networks (Kipf and Welling, 2016), while better modeling the dimension of time (i.e., chronological difference between a query and a document) is also crucial. Further on, to better deal with long documents, we plan to investigate text summarization by employing a state-of-the-art neural summarizer, e.g., BART of Lewis et al. (2020), or sentence selection techniques, e.g., rationale extraction (Lei et al., 2016; Chang et al., 2019), to find the most important sections or sentences and create shorter and more informative versions of queries/documents.

References

Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016), at the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 114–118, Berlin, Germany.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. 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.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019. A Game Theoretic Approach to Class-wise Selective Rationalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *International Conference on Learning Representations*.

Wei-Tsen Milly Chiang, Markus Hagenbuchner, and Ah Chung Tsoi. 2005. The wt10g dataset and the evolution of the web. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, page 938–939, New York, NY, USA. Association for Computing Machinery.

Charles Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the trec 2004 terabyte track. In *TREC*.

Ronan Cummins. 2016. A study of retrieval models for long documents and queries in information retrieval. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 795–805, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, abs/1810.04805.

Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling Diverse Relevance Patterns in Ad-Hoc Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 375–384, New York, NY, USA. Association for Computing Machinery.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *CoRR*, abs/1901.05287.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Manish Gupta and Michael Bendersky. 2015. Information retrieval with verbose queries. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 1121–1124, New York, NY, USA. Association for Computing Machinery.

Rupert Haigh. 2018. *Legal English*. Routledge.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058, Copenhagen, Denmark. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of coliee 2017. In *COLIEE@ ICAIL*, pages 1–8.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 177–192. Springer.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.

Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. 2016. Coliee-2016: evaluation of the competition on legal information extraction and entailment. In *International Workshop on Juris-informatics (JURISIN 2016)*.

Mi-Young Kim, Randy Goebel, and S Ken. 2015. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*.

Diederik P. Kingma and Jim Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 5th International Conference on Learning Representations*.

Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR*, abs/1609.02907.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas.

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.

Tom CW Lin. 2016. Compliance, technology, and modern finance. *Brook. J. Corp. Fin. & Com. L.*, 11:159.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Aldo Lipani, Mihai Lupu, Allan Hanbury, and Akiko Aizawa. 2015. Verboseness fission for bm25 document length normalization. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, page 385–388, New York, NY, USA. Association for Computing Machinery.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1101–1104, New York, NY, USA. Association for Computing Machinery.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. *CoRR*, abs/1809.01682.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*, Scottsdale, AZ.

Jiaul H. Paik and Douglas W. Oard. 2014. A fixed-point method for weighting terms in verbose informational queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 131–140, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. A summary of the coliee 2019 competition.

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.

S. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec3. *In Overview of the Third Text Retrieval Conference*, pages 109—-126.

Shazia Sadiq and Guido Governatori. 2015. *Managing Regulatory Compliance in Business Processes*, pages 265–288. Springer Berlin Heidelberg, Berlin, Heidelberg.

Michael Taylor, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. 2006. Optimisation methods for ranking functions with multiple parameters. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 585–593, New York, NY, USA. Association for Computing Machinery.

Peter M Tiersma. 1999. *Legal language*. University of Chicago Press.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(138).

Ellen M. Voorhees. 2005. The TREC Robust Retrieval Track. *SIGIR Forum*, 39(1):11–20.

Christopher Williams. 2007. *Tradition and change in legal English: Verbal constructions in prescriptive texts*, volume 20. Peter Lang.

A Dataset Compilation: Technical Details

In this section, we present the technical details associated with the compilation of both datasets described in the main paper. More specifically we present the procedure of creating both corpora as well as modelling the transposition relations between EU and UK entries.

A.1 EU corpus

The compilation of the EU corpus is more straightforward than its UK counterpart but involves some in-domain knowledge to filter unwanted legislation.

- We initially download the core metadata associated with each document in the EU corpus by utilizing the SPARQL endpoint of the EU Publications Office (`http://publications.europa.eu/webapi/rdf/sparql`) and the EURLEX platform (`https://eur-lex.europa.eu`), as a REST-ful API.
- Following the metadata collection, we proceed to filter out documents based on their type in order to retain only EU directives and regulations. This involves excluding corrigendums. Corrigendums introduce corrections to prior EU legislation. Usually these corrections are minimal and change single phrases such as ("In Regulation X, for: '... 4 July 2019 ...', read: '... 4 July 2015 ...'."). Thus these documents lack the context to be both classified and correlated with other documents. [16] and decisions, both of which are irrelevant to our use case. The final EU corpus contains approximately 60k entries.

A.2 UK corpus

Compiling the UK corpus is not as trivial, since the `legislation.gov.uk` API is not as evolved and we therefore have to manually crawl large parts of the database to build our corpus.

- The collected UK laws from the `legislation.gov.uk` portal form the initial corpus which includes approximately 100k documents.
- Similarly to our processing of the EU corpus, we only retain documents in specific legislation types (UK Public General Acts, UK Local Acts, UK Statutory Instruments and UK Ministerial Acts). We then eliminate laws that aim to align English legislation with the rest of the United Kingdom's, more specifically Scotland, Northern Ireland and Wales. The final UK corpus includes 52K UK entries.

A.3 EU2UK Transpositions

Transpositions are relations between entries in the EU and UK corpora which we use to define relevance for our retrieval tasks. Processing these relations is the most challenging aspect of compiling our datasets and involves several steps.

- We use the aforementioned SPARQL endpoint, to retrieve the transpositions between EU directives and the corresponding UK regulations

[16] See `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593684165879&uri=CELEX:32004L0038R(02)` as an example.

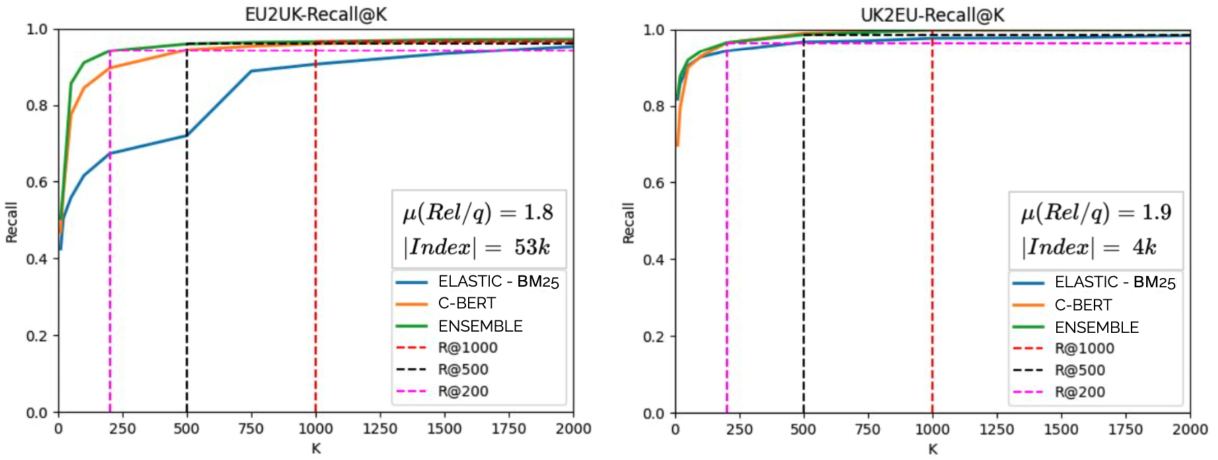

Figure 6: Recall@k, where $k \in [0, 2000]$, across the three best pre-fetchers (i.e., BM$_{25}$, C-BERT and ENSEMBLE) on the development dataset.

that implement them. We initially collect approximately 10k EU2UK pairs. In these pairs the transposed EU law is referred to by its unique portal ID but the transposing UK law is referred to by its title. This is the primary challenge in modelling the transposition relations, since mapping legislation titles to unique entries in our UK corpus is not trivial. We hypothesize that these relations are manually inserted in the database and therefore human errors make performing exact matches often impossible. Apart from the matching difficulties, some of the pairs in the pool are inserted mistakenly and hence need to be filtered.

- We first filter the noisy pairs. Pairs are considered noisy either because they are duplicates or because the do not meet some manually set criteria. In turn, duplication can occur either because identical pairs are inserted more than once or because pairs in which the UK title is mildly paraphrased are erroneously considered different. Our pool is reduced to 8k pairs after resolving the former and to 7k pairs after also resolving the latter. We further reduce the pool size by filtering pairs in which the UK title refers to non-English legislation (Scotland, Northern Ireland, Wales or Gibraltar) Non-English legislation usually has an almost identical counterpart within the pure english corpus. [17]. or in which the title does not contain certain keywords (e.g., Act, Regulation, Order, Rule). Documents that do not contain any of these keywords are not officially published in the `legislation.gov.uk` portal. Most of these are official releases from national governmental bodies, e.g. Ministries. For instance the *First Annual Report of the Inter-Departmental Ministerial Group on Human Trafficking* is not part of the UK's national legislation..
- To resolve the matching challenge, we employ a complex matching scheme where for each pair we gradually normalize the UK title until we find either a singular match or multiple ones. In the latter case, we resolve the matches with heuristics. Our normalizations include lower-casing, leading and trailing phrase removal, punctuation elimination, date removal and manually inserted substitutions.
- After reducing our pair pool and then implementing our matching scheme we can with high confidence present 4k transposition pairs which we use in our datasets.

B BERT models

All BERT variants (BERT, S-BERT, LEGAL-BERT) are publicly available from Hugging Face:

- **BERT**: The original BERT pre-trained for Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) in English Wikipedia and Books corpus. Available at `https://huggingface.co/nlpaueb/bert-base-uncased-eurlex`.
- **S-BERT**: This is the original BERT fine-tuned in STS-B NLI dataset. Available at `https://`

[17] See `https://www.legislation.gov.uk/uksi/2017/407/contents` and `https://www.legislation.gov.uk/nisr/2017/81/contents`

huggingface.co/deepset/sentence_bert.

- **LEGAL-BERT (EURLEX)**: This is the original BERT further pre-trained in EU legislaiton. Available at https://huggingface.co/nlpaueb/bert-base-uncased-eurlex.

C Selecting k for pre-fetching

In Section 4.1, we stated that we report $\mathrm{R@k}$ with $\mathrm{k} = 100$ in order to evaluate and compare pre-fetching methods. In Figure 6, we present the performance of the best pre-fetching methods (i.e., BM25, C-BERT and ENSEMBLE) for different values of $k \in [0, 2000]$ on the development set. We observe that after $k = 100$, the ENSEMBLE pre-fetcher has not significant gains in performance, thus we select $k = 100$, as a reasonable threshold.