Extreme Multi-Label Legal Text Classification: A case study in EU Legislation

Ilias Chalkidis* **Manos Fergadiotis*** **Prodromos Malakasiotis***
Nikolaos Aletras** **Ion Androutsopoulos***
* Department of Informatics, Athens University of Economics and Business, Greece
** Computer Science Department, University of Sheffield, UK
[ihalk,fergadiotis,rulller,ion]@aueb.gr, n.aletras@sheffield.ac.uk

Abstract

We consider the task of Extreme Multi-Label Text Classification (XMTC) in the legal domain. We release a new dataset of 57k legislative documents from EUR-LEX, the European Union's public document database, annotated with concepts from EUROVOC, a multidisciplinary thesaurus. The dataset is substantially larger than previous EUR-LEX datasets and suitable for XMTC, few-shot and zero-shot learning. Experimenting with several neural classifiers, we show that BIGRUs with self-attention outperform the current multi-label state-of-the-art methods, which employ label-wise attention. Replacing CNNs with BIGRUs in label-wise attention networks leads to the best overall performance.

1 Introduction

Extreme multi-label text classification (XMTC), is the task of tagging documents with relevant labels from an extremely large label set, typically containing thousands of labels (classes). Applications include building web directories (Partalas et al., 2015), labeling scientific publications with concepts from ontologies (Tsatsaronis et al., 2015), product categorization (McAuley and Leskovec, 2013), categorizing medical examinations (Mullenbach et al., 2018; Rios and Kavuluru, 2018b), and indexing legal documents (Mencia and Frnkranz, 2007). We focus on legal text processing, an emerging NLP field with many applications (Nallapati and Manning, 2008; Aletras et al., 2016; Chalkidis et al., 2017), but limited publicly available resources.

We release a new dataset, named EURLEX57K, including 57,000 English documents of EU legislation from the EUR-LEX portal. All documents have been tagged with concepts from the European Vocabulary (EUROVOC), maintained by the Publications Office of the European Union. Although EUROVOC contains more than 7,000 concepts, most of them are rarely used in practice. Consequently, they are under-represented in EURLEX57K, making the dataset also appropriate for few-shot and zero-shot learning.

Experimenting on EURLEX57K, we explore the use of various RNN-based and CNN-based neural classifiers, including the state of the art Label-Wise Attention Network of Mullenbach et al. (2018), called CNN-LWAN here. We show that both a simpler BIGRU with self-attention (Xu et al., 2015) and the Hierarchical Attention Network (HAN) of Yang et al. (2016) outperform CNN-LWAN by a wide margin. Replacing the CNN encoder of CNN-LWAN with a BIGRU, which leads to a method we call BIGRU-LWAN, further improves performance. Similar findings are observed in the zero-shot setting where Z-BIGRU-LWAN outperforms Z-CNN-LWAN.

2 Related Work

Liu et al. (2017) proposed a CNN similar to that of Kim (2014) for XMTC. They reported results on several benchmark datasets, most notably: RCV1 (Lewis et al., 2004), containing news articles; EUR-LEX (Mencia and Frnkranz, 2007), containing legal documents; Amazon-12K (McAuley and Leskovec, 2013), containing product descriptions; and Wiki-30K (Zubiaga, 2012), containing Wikipedia articles. Their proposed method outperformed both tree-based methods (e.g., FASTXML, (Prabhu and Varma, 2014)) and target-embedding methods (e.g., SLEEC (Bhatia et al., 2015), FASTTEXT (Bojanowski et al., 2016)).

RNNs with self-attention have been employed in a wide variety of NLP tasks, such as Natural Language Inference (Liu et al., 2016), Textual Entail-

Proceedings of the Natural Legal Language Processing Workshop 2019, pages 78–87
Minneapolis, Minesotta, June 7, 2019.

ment (Rocktäschel et al., 2016), and Text Classification (Zhou et al., 2016). You et al. (2018) used RNNs with self-attention in XMTC comparing with tree-based methods and deep learning approaches including vanilla LSTMs and CNNs. Their method outperformed the other approaches in three out of four XMTC datasets, demonstrating the effectiveness of attention-based RNNs.

Mullenbach et al. (2018) investigated the use of label-wise attention mechanisms in medical code prediction on the MIMIC-II and MIMIC-III datasets (Johnson et al., 2017). MIMIC-II and MIMIC-III contain over 20,000 and 47,000 documents tagged with approximately 9,000 and 5,000 ICD-9 code descriptors, respectively. Their best method, Convolutional Attention for Multi-Label Classification, called CNN-LWAN here, includes multiple attention mechanisms, one for each one of the L labels. CNN-LWAN outperformed weak baselines, namely logistic regression, vanilla BIGRUS and CNNs. Another important fact is that CNN-LWAN was found to have the best interpretability in comparison with the rest of the methods in human readers' evaluation.

Rios and Kavuluru (2018b) discuss the challenge of few-shot and zero-shot learning on the MIMIC datasets. Over 50% of all ICD-9 labels never appear in MIMIC-III, while 5,000 labels occur fewer than 10 times. The same authors proposed a new method, named Zero-Shot Attentive CNN, called Z-CNN-LWAN here, which is similar to CNN-LWAN (Mullenbach et al., 2018), but also exploits the provided ICD-9 code descriptors. The proposed Z-CNN-LWAN method was compared with prior state-of-the-art methods, including CNN-LWAN (Mullenbach et al., 2018) and MATCH-CNN (Rios and Kavuluru, 2018a), a multi-head matching CNN. While Z-CNN-LWAN did not outperform CNN-LWAN overall on MIMIC-II and MIMIC-III, it had exceptional results in few-shot and zero-shot learning, being able to identify labels with few or no instances at all in the training sets. Experimental results showed an improvement of approximately four orders of magnitude in comparison with CNN-LWAN in few-shot learning and an impressive 0.269 $R@5$ in zero-shot learning, compared to zero $R@5$ reported for the other models compared.[1] Rios and Kavuluru (2018b) also apply graph convolutions to hierarchical relations of the labels, which improves the performance on few-shot and zero-shot learning. In this work, we do not consider relations between labels and do not discuss this method further.

Note that CNN-LWAN and Z-CNN-LWAN were not compared so far with strong generic text classification baselines. Both Mullenbach et al. (2018) and Rios and Kavuluru (2018b) proposed sophisticated attention-based architectures, which intuitively are a good fit for XMTC, but they did not directly compare those models with RNNs with self-attention (You et al., 2018) or even more complex architectures, such as Hierarchical Attention Networks (HANs) (Yang et al., 2016).

3 EUROVOC & EURLEX57K

3.1 EUROVOC Thesaurus

EUROVOC is a multilingual thesaurus maintained by the Publications Office of the European Union.[2] It is used by the European Parliament, the national and regional parliaments in Europe, some national government departments, and other European organisations. The current version of EUROVOC contains more than 7,000 concepts referring to various activities of the EU and its Member States (e.g., economics, health-care, trade, etc.). It has also been used for indexing documents in systems of EU institutions, e.g., in web legislative databases, such as EUR-LEX and CELLAR. All EUROVOC concepts are represented as tuples called *descriptors*, each containing a unique numeric identifier and a (possibly) multi-word description of the concept concept, for example (1309, import), (693, citrus fruit), (192, health control), (863, Spain), (2511, agri-monetary policy).

3.2 EURLEX57K

EURLEX57K can be viewed as an improved version of the EUR-LEX dataset released by Mencia and Frnkranz (2007), which included 19,601 documents tagged with 3,993 different EUROVOC concepts. While EUR-LEX has been widely used in XMTC research, it is less than half the size of EURLEX57K and one of the smallest among XMTC benchmarks.[3] Over the past years the EUR-LEX archive has been widely expanded. EURLEX57K is a more up to date dataset including 57,000 pieces

[1] See Section 5.2 for a definition of $R@K$.

[2] https://publications.europa.eu/en/web/eu-vocabularies

[3] The most notable XMTC benchmarks can be found at http://manikvarma.org/downloads/XC/XMLRepository.html.

of EU legislation from the EUR-LEX portal.[4] All documents have been annotated by the Publications Office of EU with multiple concepts from the EUROVOC thesaurus. EURLEX57K is split in training (45,000 documents), development (6,000), and validation (6,000) subsets (see Table 1).[5]

Subset	Documents (D)	Words/D	Labels/D
Train	45,000	729	5
Dev.	6,000	714	5
Test	6,000	725	5

Table 1: Statistics of the EUR-LEX dataset.

All documents are structured in four major zones: the *header* including the title and the name of the legal body that enforced the legal act; the *recitals* that consist of references in the legal background of the decision; the *main body*, which is usually organized in articles; and the *attachments* that usually include appendices and annexes. For simplicity, we will refer to each one of *header*, *recitals*, *attachments* and each of the *main body*'s articles as *sections*. We have pre-processed all documents in order to provide the aforementioned structure.

While EUROVOC includes over 7,000 concepts (labels), only 4,271 (59.31%) of them are present in EURLEX57K. Another important fact is that most labels are under-represented; only 2,049 (47,97%) have been assigned to more than 10 documents. Such an aggressive Zipfian distribution (Figure 1) has also been noted in other domains, like medical examinations (Rios and Kavuluru, 2018b) where XMTC has been applied to index documents with concepts from medical thesauri.

The labels of EURLEX57K are divided in three categories: *frequent* labels (746), which occur in more than 50 training documents and can be found in all three subsets (training, development, test); *few-shot* labels (3,362), which appear in 1 to 50 training documents; and *zero-shot* labels (163), which appear in the development and/or test, but not in the training, documents.

4 Methods Considered

We experiment with a wide repertoire of methods including linear and non-linear neural classifiers. We also propose and conduct initial experiments

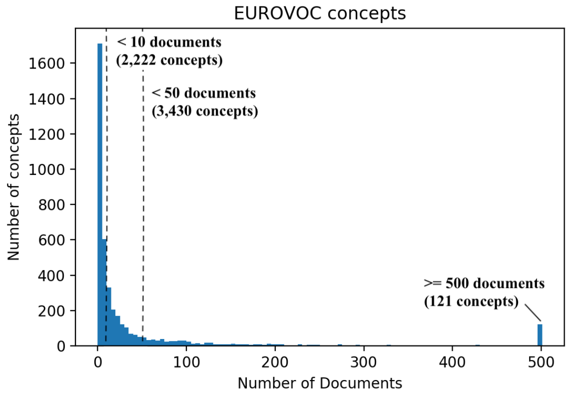


Figure 1: EUROVOC concepts frequency.

with two novel neural methods that aim to cope with the extended length of the legal documents and the information sparsity (for XMTC purposes) across the *sections* of the documents.

4.1 Baselines

4.1.1 Exact Match

To demonstrate that plain label name matching is not sufficient, our first weak baseline, Exact Match, tags documents only with labels whose descriptors appear verbatim in the documents.

4.1.2 Logistic Regression

To demonstrate the limitations of linear classifiers with bag-of-words representations, we train a Logistic Regression classifier with TF-IDF scores for the most frequent unigrams, bigrams, trigrams, 4-grams, 5-grams across all documents. Logistic regression with similar features has been widely used for multi-label classification in the past.

4.2 Neural Approaches

We present eight alternative neural methods. In the following subsections, we describe their structure consisting of five main parts:

- *word encoder* (ENC_w): turns word embeddings into context-aware embeddings,
- *section encoder* (ENC_s): turns each section (sentence) into a sentence embedding,
- *document encoder* (ENC_d): turns an entire document into a final dense representation,
- *section decoder* (DEC_s) or *document decoder* (DEC_d): maps the section or document representation to a many-hot label assignment.

All parts except for ENC_w and DEC_d are optional, i.e., they may not be present in all methods.

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

[5] Our dataset is available at http://nlp.cs.aueb.gr/software_and_datasets/EURLEX57K, with permission of reuse under European Union©, https://eur-lex.europa.eu, 1998–2019.

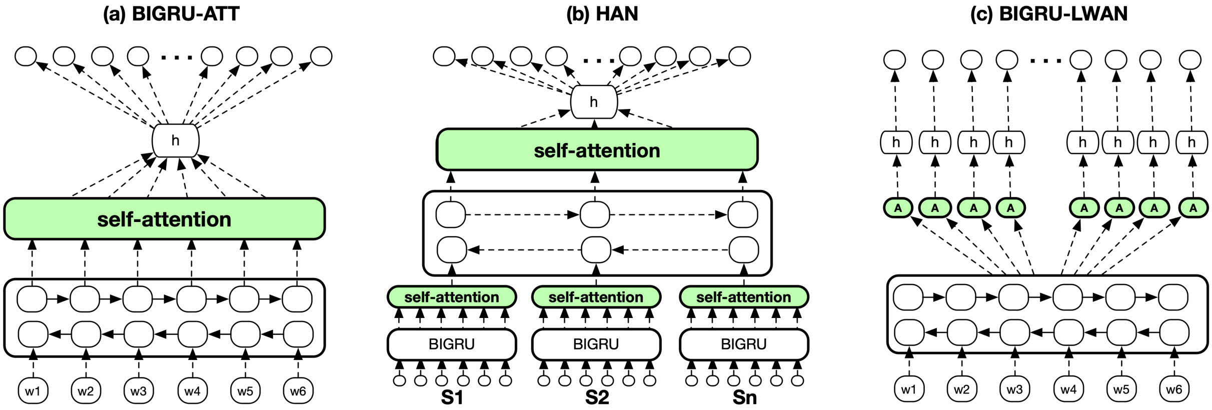


Figure 2: Illustration of (a) BIGRU-ATT, (b) HAN, and (c) BIGRU-LWAN.

4.2.1 BIGRU-ATT

In the first deep learning method, BIGRU-ATT (Figure 2a), ENC_w is a stack of BIGRUs that converts the pre-trained word embeddings (w_t) to context-aware ones (h_t). ENC_d employs a self attention mechanism to produce the final representation d of the document as a weighted sum of h_t:

$$a_t = \frac{\exp(h_t^\top u)}{\sum_j \exp(h_j^\top u)} \quad (1)$$

$$d = \frac{1}{T}\sum_{t=1}^{T} a_t h_t \quad (2)$$

T is the document's length in words, and u is a trainable vector used to compute the attention scores a_t over h_t. DEC_d is a linear layer with $L = 4,271$ output units and sigmoid (σ) activations that maps the document representation d to L probabilities, one per label.

4.2.2 HAN

The Hierarchical Attention Network (HAN) (Yang et al., 2016), exploits the structure of the documents by encoding the text in two consecutive steps (Figure 2b). First, a BIGRU (ENC_w) followed by a self-attention mechanism (ENC_s) turns the word embeddings (w_{it}) of each section s_i with T_i words into a section embedding c_i:

$$v_{it} = \tanh(W^{(s)} h_{it} + b^{(s)}) \quad (3)$$

$$a_{it}^{(s)} = \frac{\exp(v_{it}^\top u^{(s)})}{\sum_j \exp(v_{ij}^\top u^{(s)})} \quad (4)$$

$$c_i = \frac{1}{T_i}\sum_{t=1}^{T_i} a_{it}^{(s)} h_{it} \quad (5)$$

where $u^{(s)}$ is a trainable vector. Next, ENC_d, another BIGRU with self-attention, converts the section embeddings (S in total, as many as the sections) to the final document representation d:

$$v_i = \tanh(W^{(d)} c_i + b^{(d)}) \quad (6)$$

$$a_i^{(d)} = \frac{\exp(v_i^\top u^{(d)})}{\sum_j \exp(v_j^\top u^{(d)})} \quad (7)$$

$$d = \frac{1}{S}\sum_{i=1}^{S} a_i^{(d)} c_i \quad (8)$$

where $u^{(d)}$ is a trainable vector. The final decoder DEC_d of HAN is the same as in BIGRU-ATT.

4.3 MAX-HSS

Initial experiments we conducted indicated that HAN is outperformed by the shallower BIGRU-ATT. We suspected that the main reason was the fact that the section embeddings c_i that HAN's ENC_s produces contain useful information that is later degraded by HAN's ENC_d. Based on this assumption, we experimented with a novel method, named Max-Pooling over Hierarchical Attention Scorers (MAX-HSS). MAX-HSS produces section embeddings c_i in the same way as HAN, but then employs a separate DEC_s per section to produce label predictions from each section embedding c_i:

$$p_i^{(s)} = \sigma(W^{(m)} c_i + b^{(m)}) \quad (9)$$

where p_i is an L-dimensional vector containing probabilities for all labels, derived from c_i. DEC_d aggregates the predictions for the whole document with a MAXPOOL operator that extracts the highest probability per label across all sections:

$$p^{(d)} = \text{MAXPOOL}(p_1^{(s)}, \ldots, p_S^{(s)}) \quad (10)$$

Intuitively, each section tries to predict the labels relying on its content independently, and DEC_d extracts the most probable labels across sections.

4.3.1 CNN-LWAN and BIGRU-LWAN

The Label-wise Attention Network, LWAN (Mullenbach et al., 2018), also uses a self-attention mechanism, but here ENC_d employs L independent attention heads, one per label, generating L document representations $d_l = \sum_t a_{lt} h_t$ ($l = 1, \ldots, L$) from the sequence of context aware word embeddings $h_1, \ldots, h_T$ of each document d. The intuition is that each attention head focuses on possibly different aspects of $h_1, \ldots, h_T$ needed to decide if the corresponding label should be assigned to the document or not. DEC_d employs L linear layers with σ activation, each one operating on a label-wise document representation d_l to produce the probability for the corresponding label. In the original LWAN (Mullenbach et al., 2018), called CNN-LWAN here, ENC_w is a vanilla CNN. We use a modified version, BIGRU-LWAN, where ENC_w is a BIGRU (Figure 2c).

4.4 Z-CNN-LWAN and Z-BIGRU-LWAN

Following the work of Mullenbach et al. (2018), Rios and Kavuluru (2018b) designed a similar architecture in order to improve the results in documents that are classified with rare labels. In one of their models, ENC_d creates label representations, u_l, from the corresponding descriptors as follows:

$$u_l = \frac{1}{E} \sum_{e=1}^{E} w_{le} \tag{11}$$

where w_{le} is the word embedding of the e-th word in the l-th label descriptor. The label representations are then used as alternative attention vectors:

$$v_t = \tanh(W^{(z)} h_t + b^{(z)}) \tag{12}$$

$$a_{lt} = \frac{\exp(v_t^\top u_l)}{\sum_j \exp(v_j^\top u_l)} \tag{13}$$

$$d_l = \frac{1}{T} \sum_{t=1}^{T} a_{lt} h_t \tag{14}$$

where h_t are the context-aware embeddings produced by a vanilla CNN (ENC_w) operating on the document's word embeddings, a_{lt} are the attention scores conditioned on the corresponding label representation u_l, and d_l is the label-wise document representation. DEC_d also relies on label representations to produce each label's probability:

$$p_l = \sigma(u_l^\top d_l) \tag{15}$$

Note that the representations u_l of both encountered (during training) and unseen (zero-shot) labels remain unchanged, because the word embeddings w_{le} are not updated (Eq. 11). This keeps the representations of zero-shot labels close to those of encountered labels they share several descriptor words with. In turn, this helps the attention mechanism (Eq. 13) and the decoder (Eq. 15), where the label representations u_l are used, cope with unseen labels that have similar descriptors with encountered labels. As with CNN-LWAN and BIGRU-LWAN, we experiment with the original version of the model of Rios and Kavuluru (2018b), which uses a CNN ENC_w (Z-CNN-LWAN), and a version that uses a BIGRU ENC_w (Z-BIGRU-LWAN).

4.5 LW-HAN

We also propose a new method, Label-Wise Hierarchical Attention Network (LW-HAN), that combines ideas from both HAN and LWAN. For each section, LW-HAN employs an LWAN to produce L probabilities. Then, like MAX-HSS, a MAXPOOL operator extracts the highest probability per label across all sections. In effect, LW-HAN exploits the document structure to cope with the extended document length of legal documents, while employing multiple label-wise attention heads to deal with the vast and sparse label set. By contrast, MAX-HSS does not use label-wise attention.

5 Experimental Results

5.1 Experimental Setup

We implemented all methods in KERAS.[6] We used Adam (Kingma and Ba, 2015) with learning rate $1e-3$. Hyper-parameters were tuned on development data using HYPEROPT.[7] We tuned for the following hyper-parameters and ranges: ENC output units $\{200, 300, 400\}$, ENC layers $\{1, 2\}$, batch size $\{8, 12, 16\}$, dropout rate $\{0.1, 0.2, 0.3, 0.4\}$, word dropout rate $\{0.0, 0.01, 0.02\}$. For the best hyper-parameter values, we perform five runs and report mean scores on test data. For statistical significance, we take the run of each method with the best performance on development data, and perform two-tailed approximate randomization tests

[6] https://keras.io/

[7] https://github.com/hyperopt

(Dror et al., 2018) on test data. We used 200-dimensional pre-trained GLOVE embeddings (Pennington et al., 2014) in all neural methods.

5.2 Evaluation Measures

The most common evaluation measures in XMTC are recall ($R@K$), precision ($P@K$), and *nDCG* (*nDCG*$@K$) at the top K predicted labels, along with micro-averaged F-1 across all labels. Measures that macro-average over labels do not consider the number of instances per label, thus being very sensitive to infrequent labels, which are many more than frequent ones (Section 3.2). On the other hand, ranking measures, like $R@K$, $P@K$, *nDCG*$@K$, are sensitive to the choice of K. In EURLEX57K the average number of labels per document is 5.07, hence evaluating at $K = 5$ is a reasonable choice. We note that 99.4% of the dataset's documents have at most 10 gold labels.

While $R@K$ and $P@K$ are commonly used, we question their suitability for XMTC. $R@K$ leads to unfair penalization of methods when documents have more than K gold labels. Evaluating at $K = 1$ for a document with $N > 1$ gold labels returns at most $R@1 = \frac{1}{N}$, unfairly penalizing systems by not allowing them to return N labels. This is shown in Figure 3, where the green lines show that $R@K$ decreases as K decreases, because of low scores obtained for documents with more than K labels. On the other hand, $P@K$ leads to excessive penalization for documents with fewer than K gold labels. Evaluating at $K = 5$ for a document with just one gold label returns at most $P@5 = \frac{1}{5} = 0.20$, unfairly penalizing systems that retrieved all the gold labels (in this case, just one). The red lines of Figure 3 decline as K increases, because the number of documents with fewer than K gold labels increases (recall that the average number of gold labels is 5.07).

Similar concerns have led to the introduction of R-Precision and *nDCG*$@K$ in Information Retrieval (Manning et al., 2009), which we believe are also more appropriate for XMTC. Note, however, that R-Precision requires that the number of gold labels per document is known beforehand, which is not realistic in practical applications. Therefore we propose R-Precision$@K$ ($RP@K$) where K is the maximum number of retrieved labels. Both $RP@K$ and *nDCG*$@K$ adjust to the number of gold labels per document, without unfairly penalizing systems for documents with

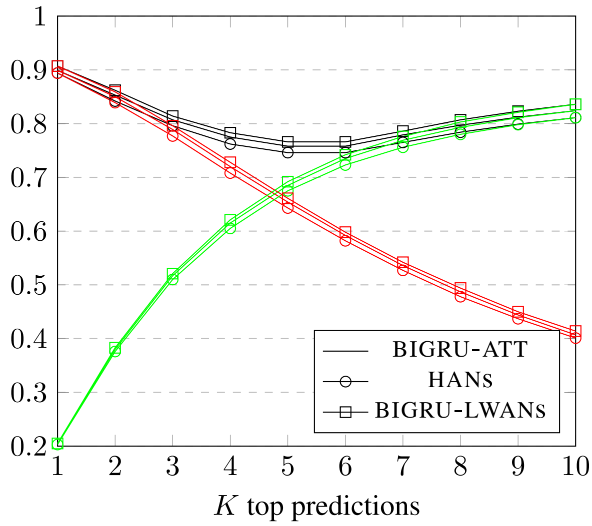


Figure 3: $R@K$ (green lines), $P@K$ (red), $RP@K$ (black) scores of the best methods (BIGRU-ATT, HANS, BIGRU-LWAN), for $K = 1$ to 10. All scores macro-averaged over test documents.

fewer than K or many more than K gold labels. They are defined as follows:

$$RP@K = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\mathrm{Rel}(n,k)}{\min(K, R_n)} \quad (16)$$

$$nDCG@K = \frac{1}{N}\sum_{n=1}^{N} Z_{Kn}\sum_{k=1}^{K}\frac{2^{\mathrm{Rel}(n,k)}-1}{\log_2(1+k)} \quad (17)$$

Here N is the number of test documents; $\mathrm{Rel}(n, k)$ is 1 if the k-th retrieved label of the n-th test document is correct, otherwise 0; R_n is the number of gold labels of the n-th test document; and Z_{Kn} is a normalization factor to ensure that *nDCG*$@K = 1$ for perfect ranking.

In effect, $RP@K$ is a macro-averaged (over test documents) version of $P@K$, but K is reduced to the number of gold labels R_n of each test document, if K exceeds R_n. Figure 3 shows $RP@K$ for the three best systems. Unlike $P@K$, $RP@K$ does not decline sharply as K increases, because it replaces K by R_n (number of gold labels) when $K > R_n$. For $K = 1$, $RP@K$ is equivalent to $P@K$, as confirmed by Fig. 3. For large values of K that almost always exceed R_n, $RP@K$ asymptotically approaches $R@K$ (macro-averaged over documents), as also confirmed by Fig. 3.

5.3 Overall Experimental Results

Table 2 reports experimental results for all methods and evaluation measures. As expected, Exact Match is vastly outperformed by machine learning

	All Labels			Frequent		Few		Zero	
	$RP@5$	$nDCG@5$	Micro-$F1$	$RP@5$	$nDCG@5$	$RP@5$	$nDCG@5$	$RP@5$	$nDCG@5$
Exact Match	0.097	0.099	0.120	0.219	0.201	0.111	0.074	0.194	0.186
Logistic Regression	0.710	0.741	0.539	0.767	0.781	0.508	0.470	0.011	0.011
BIGRU-ATT	0.758	0.789	0.689	0.799	0.813	0.631	0.580	0.040	0.027
HAN	0.746	0.778	0.680	0.789	0.805	0.597	0.544	0.051	0.034
CNN-LWAN	0.716	0.746	0.642	0.761	0.772	0.613	0.557	0.036	0.023
BIGRU-LWAN	**0.766**	**0.796**	**0.698**	**0.805**	**0.819**	**0.662**	**0.618**	0.029	0.019
Z-CNN-LWAN	0.684	0.717	0.618	0.730	0.745	0.495	0.454	0.321	0.264
Z-BIGRU-LWAN	0.718	0.752	0.652	0.764	0.780	0.561	0.510	**0.438**	**0.345**
ENSEMBLE-LWAN	**0.766**	**0.796**	**0.698**	**0.805**	**0.819**	**0.662**	**0.618**	**0.438**	**0.345**
MAX-HSS	0.737	0.773	0.671	0.784	0.803	0.463	0.443	0.039	0.028
LW-HAN	0.721	0.761	0.669	0.766	0.790	0.412	0.402	0.039	0.026

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).

methods, while Logistic Regression is also unable to cope with the complexity of XMTC.

In Section 2, we referred to the lack of previous experimental comparison between methods relying on label-wise attention and strong generic text classification baselines. Interestingly, for all, frequent, and even few-shot labels, the generic BIGRU-ATT performs better than CNN-LWAN, which was designed for XMTC. HAN also performs better than CNN-LWAN for all and frequent labels. However, replacing the CNN encoder of CNN-LWAN with a BIGRU (BIGRU-LWAN) leads to the best results overall, with the exception of zero-shot labels, indicating that the main weakness of CNN-LWAN is its vanilla CNN encoder.

5.4 Few-shot and Zero-shot Results

As noted by Rios and Kavuluru (2018b), developing reliable and robust classifiers for few-shot and zero-shot tasks is a significant challenge. Consider, for example, a test document referring to concepts that have rarely (few-shot) or never (zero-shot) occurred in training documents (e.g., 'tropical disease', which exists once in the whole dataset). A reliable classifier should be able to at least make a good guess for such rare concepts.

As shown in Table 2, BIGRU-LWAN outperforms all other methods in both frequent and few-shotlabels, but not in zero-shot labels, where Z-CNN-LWAN (Rios and Kavuluru, 2018b) provides exceptional results compared to other methods. Again, replacing the vanilla CNN of Z-CNN-LWAN with a BIGRU (Z-BIGRU-LWAN) improves performance across all label types and measures.

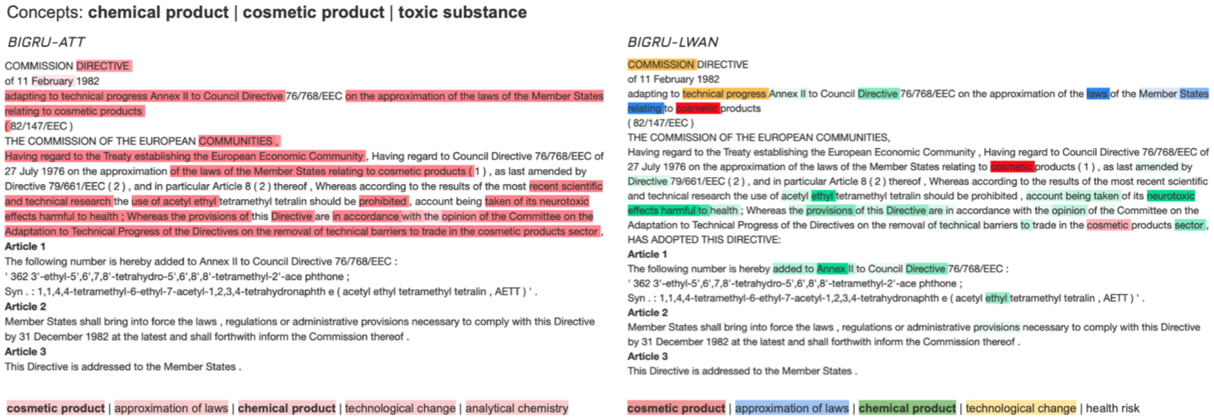
All other methods, including BIGRU-ATT, HAN, LWAN, fail to predict relevant zero-shot labels (Table 2). This behavior is not surprising, because the training objective, minimizing binary cross-entropy across all labels, largely ignores infrequent labels. The zero-shot versions of CNN-LWAN and BIGRU-LWAN outperform all other methods on zero-shot labels, in line with the findings of Rios and Kavuluru (2018b), because they exploit label descriptors, which they do not update during training (Section 4.4). Exact Match also performs better than most other methods (excluding Z-CNN-LWAN and Z-BIGRU-LWAN) on zero-shot labels, because it exploits label descriptors.

To better support all types of labels (frequent, few-shot, zero-shot), we propose an ensemble of BIGRU-LWAN and Z-BIGRU-LWAN, which outputs the predictions of BIGRU-LWAN for frequent and few-shot labels, along with the predictions of Z-BIGRU-LWAN for zero-shot labels. The ensemble's results for 'all labels' in Table 2 are the same as those of BIGRU-LWAN, because zero-shot labels are very few (163) and rare in the test set.

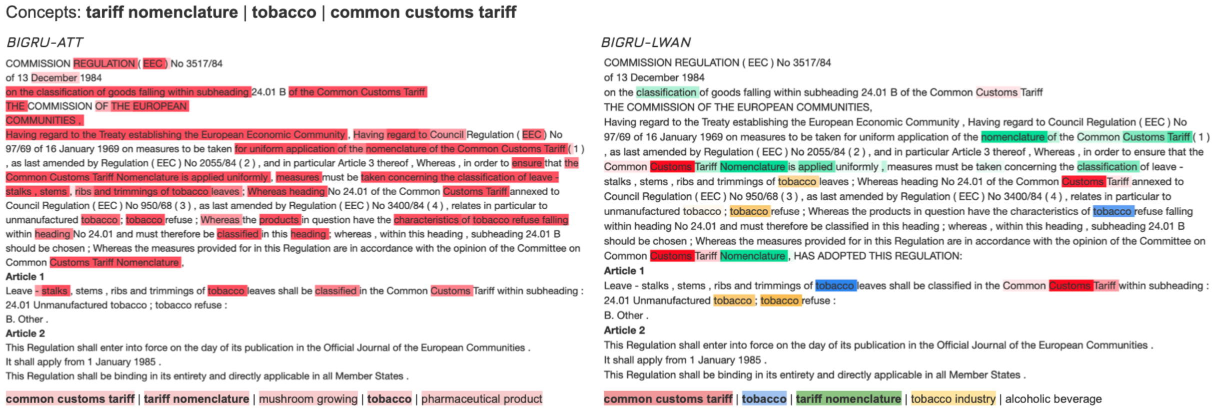
The two methods (MAX-HSS, LW-HAN) that aggregate (via MAXPOOL) predictions across sections under-perform in all types of labels, suggesting that combining predictions from individual sections is not a promising direction for XMTC.

5.5 Providing Evidence through Attention

Chalkidis and Kampas (2018) noted that self-attention does not only lead to performance improvements in legal text classification, but might also provide useful evidence for the predictions (i.e., assisting in decision-making). On the left side of Figure 4a, we demonstrate such indicative results by visualizing the attention heat-maps of BIGRU-ATT and BIGRU-LWAN. Recall that BIGRU-LWAN uses a separate attention head per label. This allows producing multi-color heat-maps (a different color per label) separately indicating which words the system attends most when predicting each label. By contrast, BIGRU-ATT uses a single attention head and, thus, the result-



(a) COMMISSION DIRECTIVE (EEC) No 82/147



(b) COMMISSION REGULATION (EEC) No 3517/84

Figure 4: Attention heat-maps for BIGRU-ATT (left) and BIGRU-LWAN (right). Gold labels (concepts) are shown at the top of each sub-figure, while the top 5 predicted labels are shown at the bottom. Correct predictions are shown in bold. BIGRU-LWAN's label-wise attentions are depicted in different colors.

ing heat-maps include only one color.

6 Conclusions and Future Work

We compared various neural methods on a new legal XMTC dataset, EURLEX57K, also investigating few-shot and zero-shot learning. We showed that BIGRU-ATT is a strong baseline for this XMTC dataset, outperforming CNN-LWAN (Mullenbach et al., 2018), which was especially designed for XMTC, but that replacing the vanilla CNN of CNN-LWAN by a BIGRU encoder (BIGRU-LWAN) leads to the best overall results, except for zero-shot labels. For the latter, the zero-shot version of CNN-LWAN of Rios and Kavuluru (2018b) produces exceptional results, compared to the other methods, and its performance improves further when its CNN is replaced by a BIGRU (Z-BIGRU-LWAN). Surprisingly HAN (Yang et al., 2016) and other hierarchical methods we considered (MAX-HSS, LW-HAN) are weaker compared to the other neural methods we experimented with, which do not consider the structure (sections) of the documents.

The best methods of this work rely on GRUs and thus are computationally expensive. The length of the documents further affects the training time of these methods. Hence, we plan to investigate the use of Transformers (Vaswani et al., 2017; Dai et al., 2019) and dilated CNNs (Kalchbrenner et al., 2017) as alternative document encoders.

Given the recent advances in transfer learning for natural language processing, we plan to experiment with pre-trained neural language models for feature extraction and fine-tuning using state-of-the-art approaches such as ELMO (Peters et al., 2018)), ULMFIT (Howard and Ruder, 2018) and BERT (Devlin et al., 2019).

Finally, we also plan to investigate further the extent to which attention heat-maps provide useful explanations of the predictions made by legal predictive models following recent work on attention explainability (Jain and Wallace, 2019).

References

Nikolaos Aletras et al. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse Local Embeddings for Extreme Multi-label Classification. In *Advances in Neural Information Processing Systems 28*, pages 730–738.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting Contract Elements. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 19–28.

Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *CoRR*, abs/1901.02860.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the Conference of the NA Chapter of the Association for Computational Linguistics*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of ACL (Long Papers)*, pages 1383–1392.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. *CoRR*, abs/1902.10186.

Alistair EW Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. 2017. MIMIC-III, a freely accessible critical care database. *Nature*.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2017. Neural Machine Translation in Linear Time. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

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

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal Machine Learning Research*, 5:361–397.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 115–124.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention. *arXiv preprint arXiv:1605.09090*, abs/1605.09090.

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

Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172.

Eneldo Loza Mencia and Johannes Frnkranz. 2007. Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Proceedings of the LWA 2007*, pages 126–132.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the NA Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1101–1111.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal Docket Classification: Where Machine Learning Stumbles. In *EMNLP*, pages 438–446.

Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artières, Georgios Paliouras, Éric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, and Patrick Gallinari. 2015. LSHTC: A Benchmark for Large-Scale Text Classification. *CoRR*, abs/1503.08581.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of NA Chapter of the Association for Computational Linguistics*.

Yashoteja Prabhu and Manik Varma. 2014. FastXML: A Fast, Accurate and Stable Tree-classifier for Extreme Multi-label Learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 263–272.

Anthony Rios and Ramakanth Kavuluru. 2018a. EMR Coding with Semi-Parametric Multi-Head Matching Networks. In *Proceedings of the 2018 Conference of the NA Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2081–2091.

Anthony Rios and Ramakanth Kavuluru. 2018b. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about Entailment with Neural Attention. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

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).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Proceedings of the 31th Annual Conference on Neural Information Processing Systems*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2048–2057.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the NA Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2018. AttentionXML: Extreme Multi-Label Text Classification with Multi-Label Attention Based Recurrent Neural Networks. *CoRR*, abs/1811.01727.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212.

Arkaitz Zubiaga. 2012. Enhancing Navigation on Wikipedia with Social Tags. *CoRR*, abs/1202.5469.