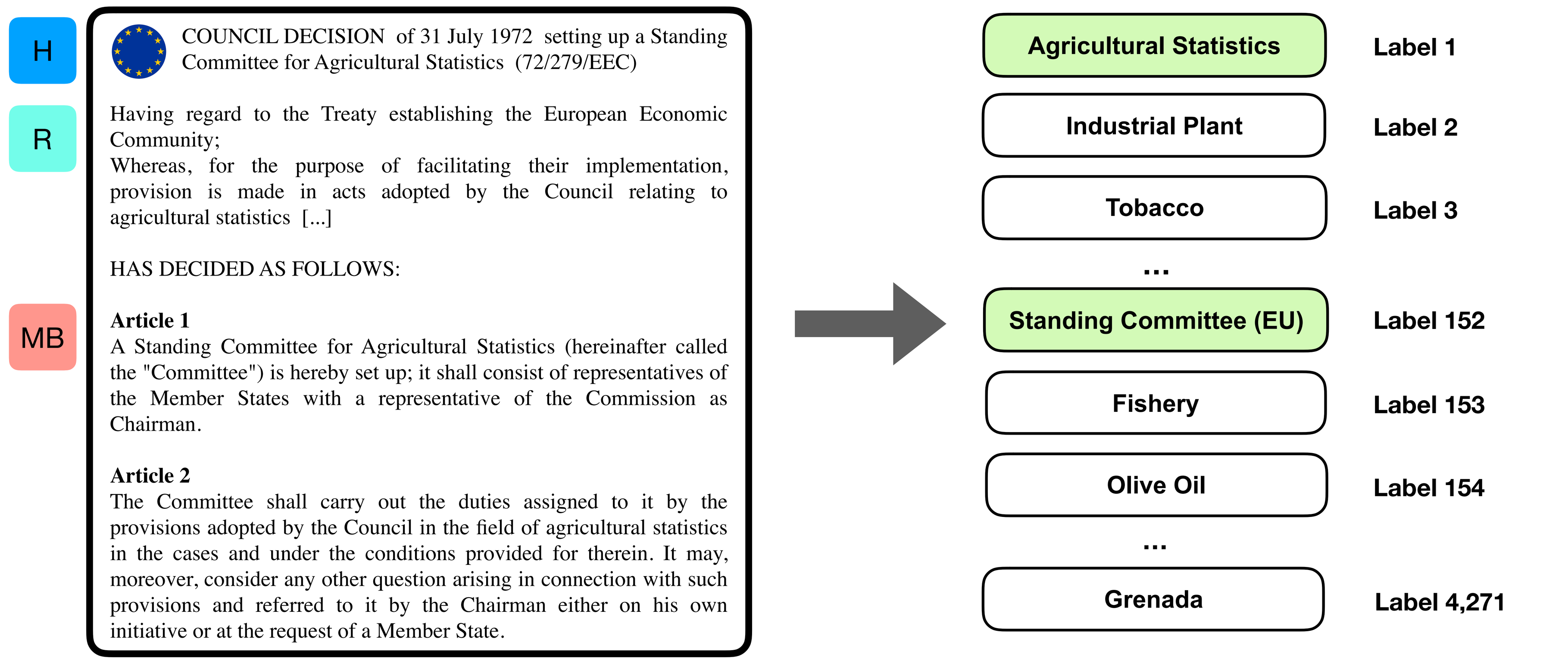
Large-Scale Multi-Label Text Classification on EU Legislation

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis and Ion Androutsopoulos

Department of Informatics, Athens University of Economics and Business, Greece

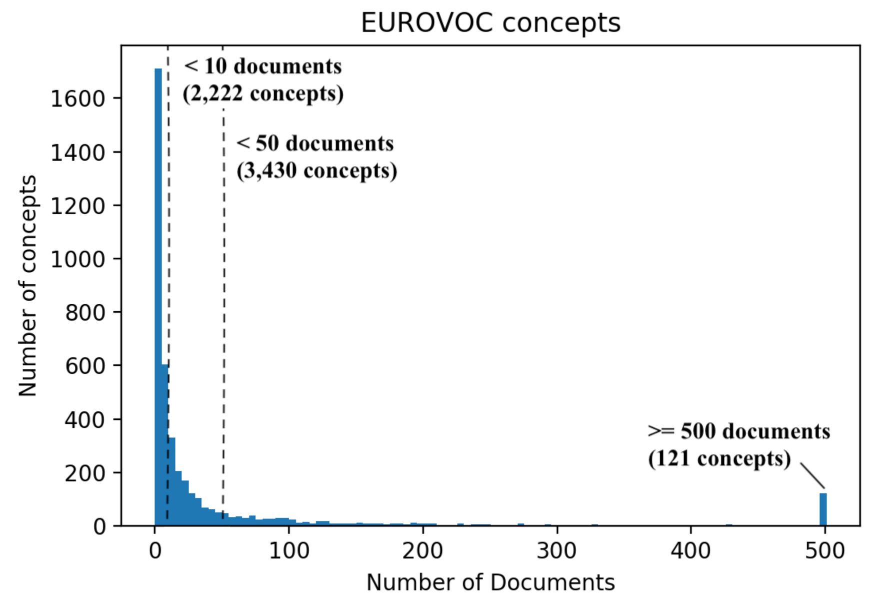
(http://nlp.cs.aueb.gr)

The Task



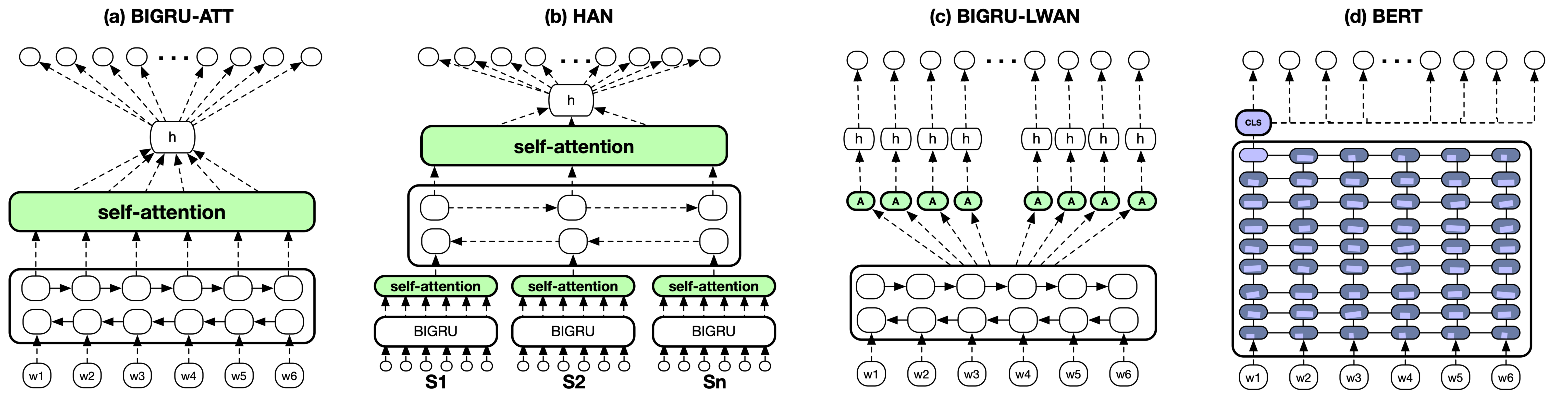
Dataset

- We release a new publicly available legal LMTC dataset, dubbed **EURLEX57K**, containing **57k English EU legislative documents** from the EUR-LEX portal, tagged with **~4.3k labels** (concepts) from the European Vocabulary (EUROVOC).
- Each EUROVOC concept is assigned with a **descriptor** (e.g., Industrial Plant, Tobacco, Spain, etc.)
- While EUROVOC includes over **7,000** concepts (labels):
  - only (59.31%) of them are present in EURLEX57K
  - only (47,97%) have been assigned >10 documents.
- Thus, we evaluate all methods for **few- and zero-shot learning**:
  - Frequent group: $D_{train} > 50$
  - Few-shot group: $1 > D_{train} >= 50$
  - Zero-shot group: $D_{train} = 0$



Methods

- **Exact Match, Logistic Regression:** A first naive baseline assigns only labels whose descriptors can be found verbatim in the document. A second one uses Logistic Regression with feature vectors containing TF-IDF scores of n-grams (n = 1,2,...,5)
- **BIGRU-ATT:** Each document is represented as the sequence of its word embeddings, which go through a stack of BIGRUs (Figure a). A document embedding (h) is computed as the sum of the resulting context-aware embeddings, weighted by the self- attention scores, and goes through a dense layer of L = 4, 271 output units with sigmoids, producing L probabilities, one per label.
- **HAN:** We use a slightly modified version of the Hierarchical Attention Network (Yang et al., 2016), where a BIGRU with self-attention reads the words of each section, as in BIGRU-ATT but separately per section, producing section embeddings. A second-level BIGRU with self-attention reads the section embeddings, producing a single document embedding (h) that goes through a similar output layer as in BIGRU-ATT (Figure b).
- **LWAN:** Unlike BIGRU-ATT, LWAN uses L independent attention heads, one per label, generating L document embeddings from the sequence of context-aware embeddings produced by a CNN or BIGRU encoder, respectively. Each document embedding (h) is specialized to predict the corresponding label and goes through a separate dense layer with a sigmoid, to produce the probability of the corresponding label (Figure c).
- **ZERO-LWAN:** Rios and Kavuluru (2018) designed a model similar to LWAN to deal with rare labels. In ZERO-LWAN, the attention scores and the label probabilities are produced by comparing the context-aware embeddings that the CNN or BIGRU encoder produces and the label-specific document embeddings (h), respectively, to label embeddings. Each label embedding is the centroid of the pre-trained word embeddings of the label's descriptor. By contrast, LWAN does not consider the descriptors of the labels.
- **BERT:** For a new target task, a task-specific layer is added on top of BERT. The extra layer is trained jointly with BERT by fine-tuning on task-specific data. We add a dense layer on top of BERT, with sigmoids, that produces a probability per label (Figure d). Unfortunately, BERT can currently process texts up to 512 word-pieces, which is too small for the documents of EURLEX57K. Hence, BERT can only be applied to truncated versions of our documents.



Experimental Results

	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
ZERO-CNN-LWAN	0.684	0.717	0.618	0.730	0.745	0.495	0.454	0.321	0.264
ZERO-BIGRU-LWAN	0.718	0.752	0.652	0.764	0.780	0.561	0.510	**0.438**	**0.345**
BIGRU-LWAN (L2V)	0.775	0.804	0.711	0.815	0.828	0.656	0.612	0.034	0.024
BIGRU-LWAN (L2V) *	0.770	0.796	0.709	0.811	0.825	0.641	0.600	0.047	0.030
BIGRU-LWAN (ELMO) *	0.781	0.811	0.719	0.821	0.835	0.668	0.619	0.044	0.028
BERT-BASE *	**0.796**	**0.823**	**0.732**	**0.835**	**0.846**	**0.686**	**0.636**	0.028	0.023

Alternative Word Representations

	RP@5	nDCG@5	Micro-F1
GLOVE	0.766	0.796	0.698
LAW2VEC	0.775	0.804	0.711
GLOVE + ELMO	0.777	0.808	0.714
LAW2VEC + ELMO	**0.781**	**0.811**	**0.719**

Using Particular Document Zones

	μwords	RP@5	nDCG@5	Micro-F1
HEADER	43	0.747	0.782	0.688
RECITALS	317	0.734	0.765	0.669
HEADER + RECITALS	360	0.765	0.796	0.701
MAIN BODY	187	0.643	0.674	0.590
FULL TEXT	727	**0.766**	**0.797**	**0.702**

Other Recent Publications

(nlp.cs.aueb.gr/publications.html)

- I. Chalkidis and I. Androutsopoulos, "A Deep Learning Approach to Contract Element Extraction". Proceedings of the *30th International Conference on Legal Knowledge and Information Systems (JURIX 2017)*, Luxembourg, pp. 155-164, 2017.
- I. Chalkidis, I. Androutsopoulos and A. Michos, "Obligation and Prohibition Extraction Using Hierarchical RNNs". Proceedings of the *56th Annual Meeting of the Association for Computational Linguistics* (ACL 2018), Melbourne, Australia, pp. 254-259 (short papers), 2018.
- I. Chalkidis, I. Androutsopoulos and N. Aletras, "Neural Legal Judgment in English". Proceedings of the *57th Annual Meeting of the Association for Computational Linguistics* (ACL 2019), Florence, Italy, (short papers), 2019.

Resources

Dataset: **http://nlp.cs.aueb.gr/software_and_datasets/EURLEX57K**

Code: **https://github.com/iliaschalkidis/lmtc-eurlex57k**