Information retrieval for AI-assisted legislative drafting

Michele Corazza, Monica Palmirani

University of Bologna

HYPERMODELEX

Motivation

- The usage of generative models in legislative drafting is only in its infancy, but we already see some examples:

- However, the hallucinations that are produced by LLMs are problematic, as they can output text that is well-formed, which in reality contains factually incorrect information.

Objective

- The creation of a LLM for the legislative drafting in the context of European institutions; in particular, the model should be able to draft the preambles of European legislative documents;
- it is necessary for the resulting LLM to incorporate information about other documents that need to be referenced in the preamble;
- this presentation is concerned the challenges that are concerned when creating an information retrieval system for the references, as well as some possible solutions.

Challenges and considerations

- The structured nature of legislative documents;
- Ad-hoc or pre-trained models;
- The length of legislative documents;
- The legislative system as a network

The structured nature of legislative documents: example

Article 5

Principles relating to processing of personal data

1. Personal data shall be:

(a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');

(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').

2. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability').

The structured nature of legislative documents

- It is desirable to use an approach that denotes the **structure** and **semantic components** of legislative documents;

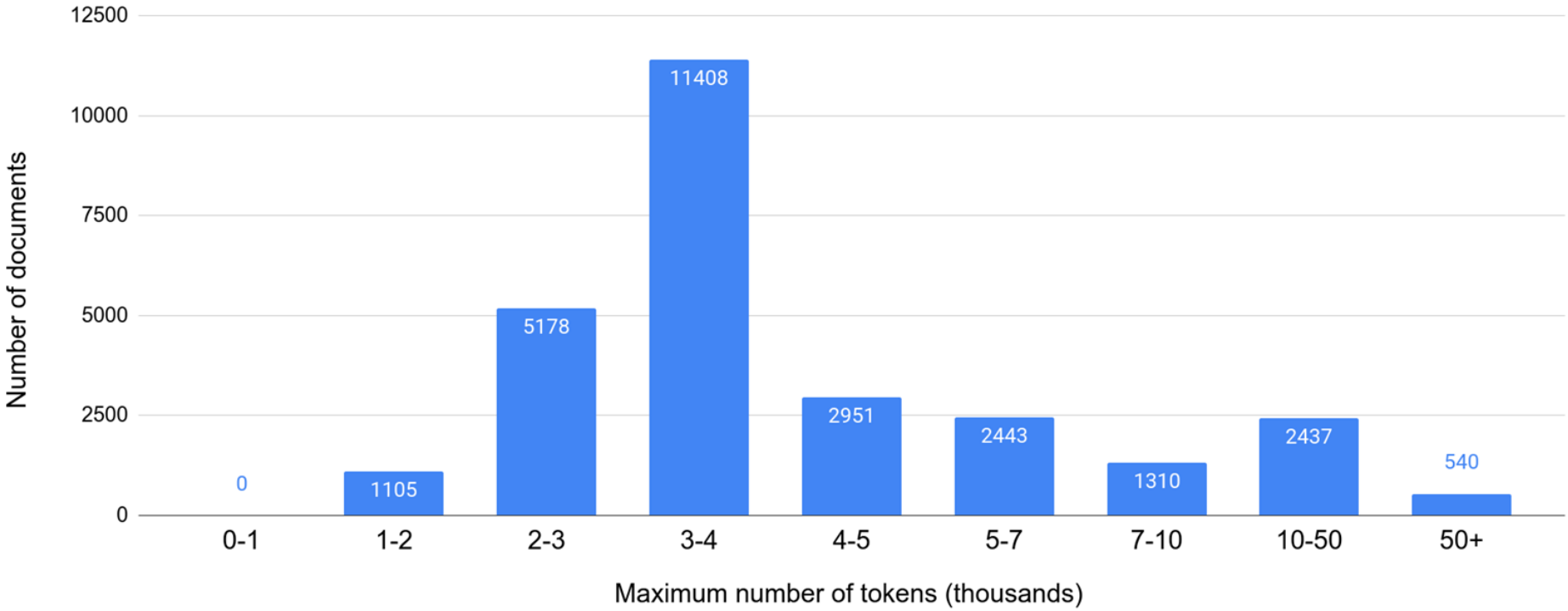
- Akoma Ntoso then is the logical solution, as it is able to annotate the structural nature of legislative documents;
- Additionally, it annotates the **references** and **temporal** information.

Ad-hoc and pre-trained models: trade-offs

- Using pre-trained models (BERT, RoBERTa) that can be fine-tuned for the domain reduces amount of in-domain data required;
- However, this leads to the concrete possibility that the model was trained on legislative documents from other legal traditions (eg civil law vs common law);
- Additionally, the model might have been trained on the documents that will be used to evaluate it (*data contamination*);
- Finally, our choice to use Akoma Ntoso documents as inputs to the model could lead to problems for the pre-trained tokenizers, especially those using Byte Pair Encoding.

For these reasons, we argue for the creation of an ad-hoc model trained only on European legislative documents.

The length of legislative documents



The length of legislative documents

- In order to process the long documents such as the legislative documents of the European Union, it is possible to use **hierarchical models**;
- An approach based on Hierarchical BERT [1] can be used, and it would allow the model to process a maximum of 512*512=262144 tokens;
- While Hierarchical BERT uses sentences to split the document, using Akoma Ntoso it is possible to traverse the XML tree and split the document using its structural components (articles, paragraphs, etc).

[1] Lu, J., Henchion, M., Bacher, I., Namee, B.M.: A sentence-level hierarchical bert model for document classification with limited labelled data. Proceedings of the International Conference on Discovery Science (2021)

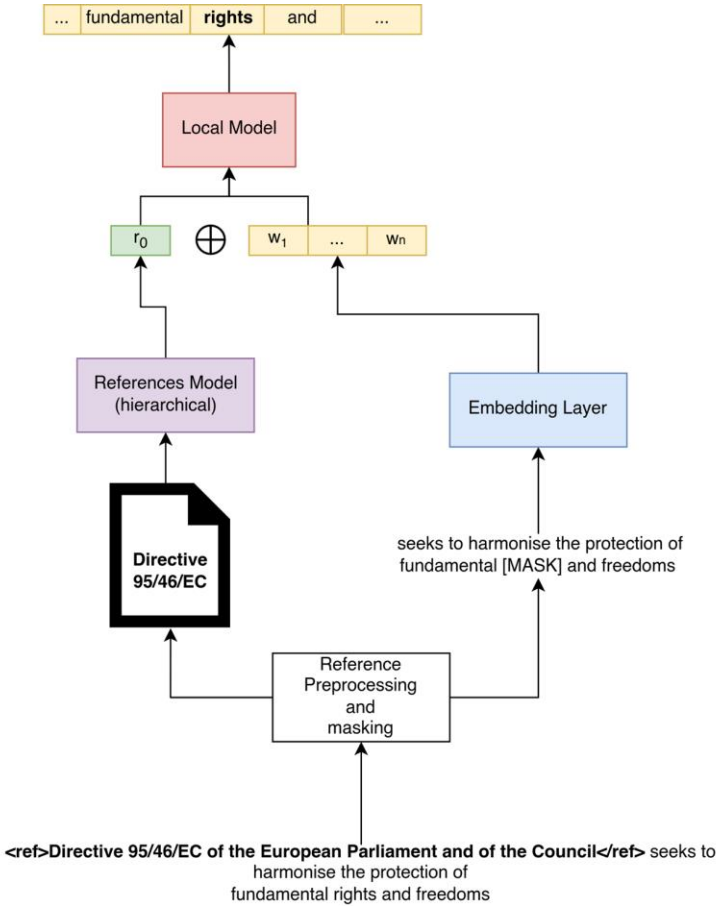
The legislative system as a network

- Another crucial aspect is the extensive usage of normative references to other documents;
- The legislative system, then, is a complex network of documents liked by hypertextual citations;
- Sometimes, they are used to reference **definitions**, so any model operating on legislative documents should be able to consider their content.

> (45) 'personal data' means personal data as defined in Article 4, point (1), of Regulation (EU) 2016/679;

(Source: Regulation (EU) 2023/1114)

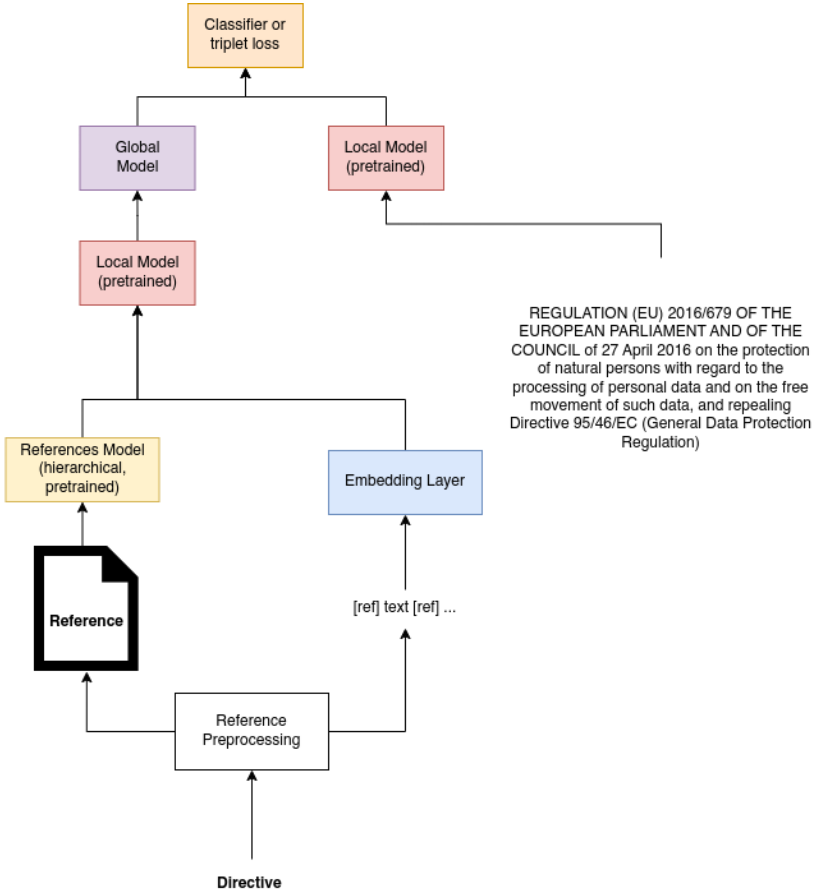
Pre-training: MLM and references model



Scalability

- An information retrieval system for normative references should be able to operate on **hundreds of thousands of documents**;
- This poses a challenge of scale, as it is not feasible to apply a deep neural model to hundreds of thousands of document each time the system is used;
- In order to mitigate this issue, our proposal is to pre-compute a vector representing each document, so that **triplet loss** can be used, allowing the application of a distance metric to the vectors (eg Euclidean distance);
- The vectors are then saved, so that the larger transformer-based portion of the model is applied only once to each document.

Fine-tuning


Classifier or triplet loss
Global Model
Local Model (pretrained)
Local Model (pretrained)
REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
References Model (hierarchical, pretrained)
Embedding Layer
Reference
[ref] text [ref] ...
Reference Preprocessing
Directive

Conclusions

The preliminary analysis of this presentation is a necessary step for the creation of an information retrieval system for legislative documents. In particular, we addressed the following challenges:

- The structural nature of legislative documents;
- The length of legislative documents;
- The trade-offs between an ad-hoc model and a pre-trained one;
- The scalability of the information retrieval system