Artificial Intelligence and Law
https://doi.org/10.1007/s10506-023-09387-2

ORIGINAL RESEARCH

DiscoLQA: zero-shot discourse-based legal question answering on European Legislation

Francesco Sovrano[1,2] · **Monica Palmirani**[3] · **Salvatore Sapienza**[3] · **Vittoria Pistone**[3]

Accepted: 1 December 2023

Abstract

The structures of discourse used by legal and ordinary languages share differences that foster technical issues when applying or fine-tuning general-purpose language models for open-domain question answering on legal resources. For example, longer sentences may be preferred in European laws (i.e., Brussels I bis Regulation EU 1215/2012) to reduce potential ambiguities and improve comprehensibility, distracting a language model trained on ordinary English. In this article, we investigate some mechanisms to isolate and capture the discursive patterns of legalese in order to perform zero-shot question answering, i.e., without training on legal documents. Specifically, we use pre-trained open-domain answer retrieval systems and study what happens when changing the type of information to consider for retrieval. Indeed, by selecting only the important parts of discourse (e.g., elementary units of discourse, EDU for short, or abstract representations of meaning, AMR for short), we should be able to help the answer retriever identify the elements of interest. Hence, with this paper, we publish Q4EU, a new evaluation dataset that includes more than 70 questions and 200 answers on 6 different European norms, and study what happens to a baseline system when only EDUs or AMRs are used during information retrieval. Our results show that the versions using EDUs are overall the best, leading to state-of-the-art F1, precision, NDCG and MRR scores.

Keywords Legal question answering · European Legislation · Knowledge graph extraction · Discourse theory · Abstract meaning representations · Private international law · European arrest warrant · GDPR · Electronic signature

1 Introduction

We are witnessing a growing need for the digitisation of our society, which requires great interdisciplinary efforts in law, information technology and engineering. This need has led to the birth of institutions such as the Ministry of Digital Governance

Extended author information available on the last page of the article

Published online: 10 January 2024

of Greece and the Australian Digital Transition Agency, or long-term plans like the European Digital Transition Action Plan and many others.

In the literature of AI, answering questions using an extensive collection of documents of diversified topics (i.e., Private International Law) is called open-domain Question Answering (QA). Modern open-domain QA systems usually combine traditional information retrieval techniques and neural reading comprehension models. Nevertheless, neural reading comprehension of legal texts (e.g., European legislation) is challenging because legalese is rarer, mercurial and in many ways different from a commonly used natural language. Hence, the difference between legal and ordinary languages does foster technical issues when applying or fine-tuning general-purpose language models for open-domain question answering on legal resources. This is especially true when the meaning of a legal document is encoded in its (discourse) structure in a way that is different from the spoken language. For example, long sentences or more "formal" writing may be preferred in legislative documents (e.g., Brussels I bis Regulation EU 1215/2012) to reduce potential ambiguities and improve comprehensibility. However, the noise introduced by the excessive length of the sentence or their unusual structure can distract a language model trained in ordinary English, pushing it to commit more errors.

As a result, standard neural reading comprehension models may only be able to represent the semantics of a legal text if they are adequately specialised to do it. This is because legalese is not repetitive. It is canonical and has semantic terminology that tends to avoid polysemy and to be used punctually in particular contexts as if the sentences it forms were governed by formal rules. Hence, applying these formal rules impacts the discourse structure, as suggested by Sovrano et al. (2022).

Here, we expand the work published by Sovrano et al. (2020), investigating some mechanisms to perform "zero-shot" legal question answering. More specifically, "zero-shot" means that question answering is performed through pre-trained language models (e.g., a model that is trained on generic non-legal documents) without fine-tuning them on the downstream legal task of question answering. In this sense, zero-shot legal question answering can be a necessary solution for all those tasks characterised by a paucity of data (e.g., European hard laws, the resolutions of the United Nations General Assembly) and for which we want to train AI-based solutions through machine learning without having enough information for effective fine-tuning. Conversely, zero-shot legal question answering might be less helpful whenever data are abundant (e.g., American case law or privacy policies).

In this article, we investigate the role of discourse structure in legalese, trying to understand and exploit its importance in encoding the meaning of legal documents. The goal of this investigation is also practical, not just theoretical. Understanding how legalese differs from its spoken counterpart can help solve the data scarcity problem in legalese processing/comprehension. This would allow us to better exploit generic language models not calibrated to a downstream legal task or even not trained on legal documents, as shown throughout the paper.

Specifically, we use open-domain QA systems based on information retrieval and neural reading comprehension and study what happens when changing the type of information to consider for retrieval. These QA systems encode all the possible answers (e.g., parts of articles, recitals) with a general-purpose neural

model and then use the encoding for fast similarity-based retrieval. Usually, these answers are just a short part (a grammatical sub-tree) of one sentence or paragraph, especially if the whole is very long. Suppose the neural model is not specialised in legalese. In that case, it will likely fail to identify and capture the importance of grammatical sub-trees that are uncommon in the spoken language. Hence, by selecting only those grammatical sub-trees deemed the most important, we should be able to help the information retriever and the QA system by partially hiding noise within answers. To identify these important grammatical sub-trees, we used the theory of Elementary Discourse Units (EDUs) (Prasad et al. 2008) and the theory of Abstract Meaning Representations (AMRs) (Banarescu et al. 2013).

In other words, we show how to produce more effective answer retrieval tools by capturing discourse structure, leveraging existing tools for QA specialised in common natural languages. Therefore, to shed more (empirical) light on what constitutes *meaning* in legalese, we decided to design an experiment focused on understanding whether there is a benefit in using only EDUs or AMRs, as triplets in the knowledge graphs extracted by the pipeline proposed by Sovrano et al. (2020). We devised a simple experiment where we study what happens to the baseline QA system when using EDUs or AMRs during information retrieval.

In particular, to evaluate our results, we present a new dataset called Q4EU that extends Q4PIL (Sovrano et al. 2021) with 3 European norms, for a total of 72 unique questions and 225 expected answers (in the form of articles and recitals) on 6 heterogeneous European norms spanning from Private International Law to Human Rights Law (i.e., the General Data Protection Regulation, UE 2016/679), from regulations of electronic signatures to the European arrest warrant.

The results of our experiments show that the versions using EDUs are overall the best, leading to state-of-the-art top-k precision and F1 scores for all the values of k we considered. Our instances of DiscoLQA were able to generalize across the different legal sub-domains tested, even if the deep language models involved were not pre-trained on legal corpora.

However, we tested and evaluated DiscoLQA on specific European norms and a relatively small dataset, without using deep language models pre-trained on legal corpora.

Our contribution is threefold:

1. We show where a general-purpose language model may fail when applied to legal documents, hinting at how to intervene for effective fine-tuning or re-training. In other words, we show that legalese's semantics may be encoded differently. Identifying the sources of meaning may be beneficial for effectively improving the state-of-the-art neural reading comprehension of legal documents.
2. We show a way to effectively use discourse analysis for legal question answering, improving state-of-the-art without fine-tuning or re-training the language models on the regulations at hand.
3. We publish Q4EU, a new evaluation dataset for legal question answer retrieval that extends the work by Sovrano et al. (2021).

For reproducibility purposes, we also publish on GitHub the source code of DiscoLQA.[1]

This paper is structured as follows. In Sect. 2, we discuss the related work on (legal) QA, while in Sect. 3 we give all the necessary background information to understand the pipeline of algorithms presented in Sect. 4. Finally in Sect. 5 we discuss our experiment and present the Q4EU dataset, analysing the results in Sect. 7 while pointing to future work in Sect. 8.

2 Related work

Legal QA is a relatively recent field of study in the context of AI and Law, with many exciting solutions available today. Some of these solutions follow end-to-end approaches, exploiting existing language models. In contrast, some others try to exploit ontologies and knowledge graphs by framing QA as a task of information retrieval.

On the one hand, we can see a paucity of end-to-end generic solutions to legal QA that are usually focused only on particular and narrow applications for which large enough datasets are available. Instead, when no large dataset is available for training, we generally have that using deep language models pre-trained on ordinary English does not always produce good results. Zheng et al. (2021) showed that the more complex the legal reasoning task to answer, the less effective the fine-tuning could be. An example of end-to-end QA system is the work by Kim et al. (2015), where a deep neural network is trained on a dataset of Boolean questions from Japanese legal bar exams. Another interesting example is the work by Ravichander et al. (2019), proposing an end-to-end question-answering solution for privacy policies.

On the other hand, an example of an answer retrieval system specific to Private International Law is the one proposed by Sovrano et al. (2020). It consists of a combination of TF-IDF and some deep language models to retrieve pertinent answers from an automatically extracted knowledge graph of contextualised grammatical sub-trees. In particular, the knowledge graph is aligned to a legal ontology based on Ontology Design Patterns (i.e., agent, role, event, temporal parameter, action) to mirror the legal significance of the relationships within and among the provisions. In this sense, we extend the work by Sovrano et al. (2020), trying to overcome some of the issues of using language models not trained in legalese.

While another example of an answer retrieval system is the work by Vold and Conrad (2021), comparing the performance of a deep learning-based solution with that of a traditional SVM. In particular, Vold and Conrad (2021) fine-tuned a deep language model (called RoBERTa) on a dataset of questions about privacy policies (that usually use a language closer to spoken English rather than legalese), obtaining better results than with an SVM.

[1] https://github.com/Francesco-Sovrano/DiscoLQA

3 Background

In this section, we provide all the necessary background information to understand state-of-the-art automated question-answering and the relationship between discourse theory and legalese.

3.1 Question answering and law

Natural language processing/understanding is of utmost importance in the intersection of AI and Law. This is why many works in this field have focused on general-purpose state-of-the-art language models for the generation of word/sentence embeddings (Shao et al. 2020; Condevaux et al. 2019; Vink et al. 2020).

For example, Bommarito et al. (2018) published a framework for natural language processing and information extraction for legal and regulatory texts. While Chalkidis and Kampas (2019) proposed one of the first models for legal word embeddings. Also, the Incorporated Council of Law Reporting for England and Wales (ICLR 2019) published Blackstone, a library meant to allow researchers and engineers to automatically extract information from long, unstructured legal texts (such as judgements, skeleton arguments, scientific articles, Law Commission reports, pleadings). More generally, natural language processing for legal texts has recently raised a lot of interest, highlighting "the need to create a bridge between conceptual questions, such as the role of legal interpretation in mining and reasoning, as well as computational and engineering challenges, such as the handling of big legal data and the complexity of regulatory compliance" (Robaldo et al. 2019).

Automating *legal reasoning* is not a trivial task, as it requires a deep understanding of language, non-monotonic logic and the theory of interpretation, as well as sufficient flexibility to handle the plethora of changes to which law and hermeneutics are subject over time. Current state-of-the-art AI for reasoning is divided into two approaches: the symbolic and the sub-symbolic. The symbolic approach draws from formal languages and logic. It requires every component of the reasoning to be an abstract symbol with a pre-defined and context-independent interpretation of its meaning, making the AI based on this approach hardly compatible with natural languages such as English, Chinese, and Spanish. On the other hand, the sub-symbolic approach draws from recent advancements in deep learning. Exploiting large amounts of data, it can "understand" natural language and visual inputs in a scalable and highly effective way. However, it loses transparency by working on non-symbolic representations (i.e., arbitrary numerical vectors) frequently not interpretable.

Non-monotonic reasoners based on Defeasible Logic (Lam and Governatori 2009), Deontic Logic (Hage 2000) and Argumentation (Gordon and Walton 2009) are famous examples of symbolic AI applied to the legal domain. All require legal documents to be translated (manually) from their original natural language into some particular formal language upon which classical logical reasoning can be applied. This type of reasoner usually struggles to scale to handle natural language (i.e., English) inputs such as documents and questions.

On the other hand, the sub-symbolic approach is more versatile and well-known to be more easily applied directly to natural language documents. Famous sub-symbolic approaches to (legal) reasoning are the so-called QA algorithms. As suggested by Xie et al. (2020); Cao et al. (2019); Zhang et al. (2018); Hudson and Manning (2019) and others, in many cases, question answering can be seen as an instance of reasoning. These QA algorithms are usually trained end-to-end to extract short (i.e., 2–3 words) answers from a whole document (text or image) to match a given question.

The most common end-to-end QA algorithms, i.e. those collected by Wolf et al. (2020), rely on Transformers (Vaswani et al. 2017). Hence they have quadratic complexity in the size of the whole document to be searched for an answer. This characteristic makes end-to-end QA based on Transformers fail in all those situations where collections of large documents of diversified topics (i.e., Private International Law) are involved, or parts of the same answer are scattered across multiple documents. A solution to this problem is seen in *Question-Answer Retrieval*, also known as *Dense Passage Retrieval* or open-domain QA (Chen and Yih 2020). Modern open-domain QA systems usually combine traditional information retrieval techniques and neural reading comprehension models. These QA systems encode all the identified possible answers (e.g., parts of articles, recitals) with a general-purpose neural model. Then they use the encoding for fast similarity-based retrieval. Therefore, differently from end-to-end QA, *Question-Answer Retrieval* is less end-to-end, requiring the *a priori* identification of the possible snippets of text functioning as answers, but it is much faster. In fact, it has a complexity that is usually proportional to the product of the size of the context (normally a small paragraph) and the size of the answer (commonly smaller than the context).

Among the most important *Question-Answer Retrieval* models, we distinguish between those that use the answer's context for the generation of embeddings[2] (Yang et al. 2020; Karpukhin et al. 2020; Roy et al. 2020) and those who do not (Chen et al. 2020).

3.2 Discourse theory and legal language

The relation between discourse theory and legalese is complicated and still open to discussion. Discourse theory is a branch of linguistics that studies how coherence and cohesive relations can be the threads that make up a text to form a discourse. A discourse is said to be coherent if all of its pieces belong together, while it is said to be cohesive if its elements have some common thread. Sanders et al. (1992) identified two *requirements for a theory of discourse*:

- *Descriptive adequacy*: A theory discourse structure makes it possible to describe the structure of all kinds of (natural) texts.

[2] Intuitively, using the answer's context should help the answer embedder to contextualise and disambiguate better, producing more high-quality embeddings.

- *Psychological plausibility*: A theory of discourse structure should at least generate plausible hypotheses on the role of discourse structure in constructing cognitive representation.

In recent years, many different theories of discourse have been spelt out, each with different pros and cons. Among them, we cite the Rhetorical Structure Theory (Mann and Thompson 1988), assuming that discourse is structured as a tree, the Segmented Discourse Representation Theory (Lascarides and Asher 2007) assuming that discourse is structured as a graph (therefore allowing long-distance attachments), and the theory of EDUs (Miltsakaki et al. 2004; Prasad et al. 2008; Webber et al. 2019) making no assumption on the text structure. Common to them is probably the identification of something that may be called Elementary Discourse Unit (EDU). EDUs are spans of text denoting a single event serving as a complete, distinct unit of information that the surrounding discourse may connect to Stede (2013). EDUs can be combined to form many different types of discourse Fludernik (2000); D'Angelo (1984) including: argumentation, exposition, description, narration.

The theory of EDUs encoded by the Penn Discourse Treebank (PDTB) model is considered one of the most generic theories of discourse. Indeed, PDTB is data-driven (based on lexically grounded relations) and makes little assumptions about the underlying language. As a result, with little or no change in annotative style, PDTB appears to be usable for modelling discourses of natural languages belonging to different families (Zufferey and Degand 2017), e.g., Chinese, Arabic, and Hindi. In particular, PDTB is based on the assumption that "the meaning and coherence of a discourse result partly from how its constituents relate to each other". Therefore *discourse relations* are defined as semantic relations between abstract objects (or EDUs) mentioned in discourse and connected by explicit (e.g., "but", "then", "for example", and "although") or implicit relations. According to PDTB, discourse relations can be of one of 4 main types: temporal, contingency (causality, purpose, etc.), expansion, and comparison. PDTB-style annotations and the other theories of discourse have inspired an ISO standard (Prasad and Bunt 2015).

The application of PDTB to legalese has been explored by some Robaldo et al. (2008); Cabrio et al. (2013), but has yet to have much follow-up. The point is that ordinary discourse theory is better suited to judgments, Hansard reports, testimonies and reports of debates. Instead, it seems unsuited to legislative texts and contracts, for which a specific vocabulary (e.g., definitions) or textual structure (e.g., hierarchy) is used to identify meaning through interpretation theory. Indeed, legislative texts have a deeper structure than common sentences. For example, a list has a legal meaning of conditions linked together by specific semantics. Furthermore, the classical linguistic structures based on discourse connectives tend to be used differently in law. Legal connectives do not have the same semantic value as everyday discourse. They are operators of deontic rules with multiple meanings (e.g., "xor", "or", "and"). Also, some discourse structures tend not to be used at all because they are not a good practice in legal drafting (e.g., "but" and "for example").

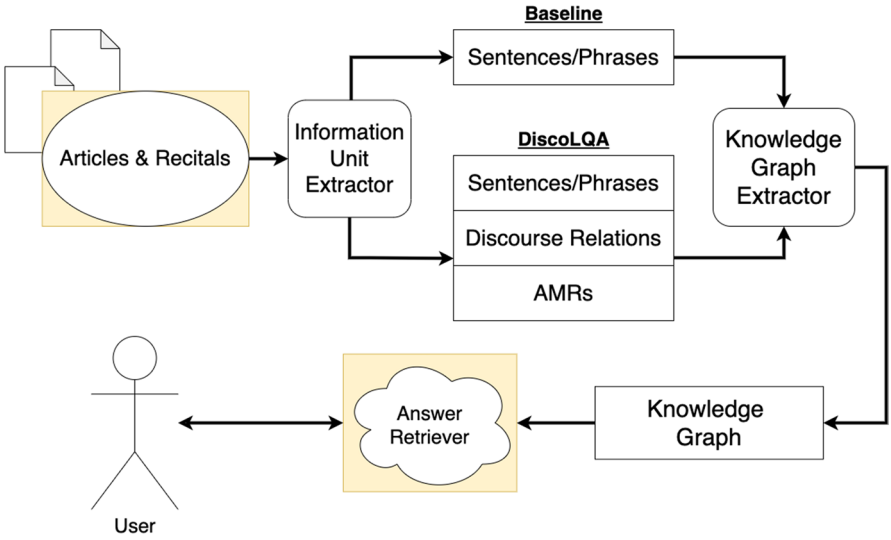

Fig. 1 Sketch of the pipeline used in the *baseline* and DiscoLQA. The *baseline* extracts only clauses from the source texts (articles, recitals, commission statements, etc.). DiscoLQA also extracts discourse relations and AMR as *information units*. The *information units* are then passed to the knowledge graph extractor that produces a graph used by the *Question-Answer Retriever*

4 DiscoLQA: discourse theory for legal question answering

This paper proposes a novel pipeline of algorithms called DiscoLQA, short for Discourse-based Legal Question Answering. DiscoLQA is based on the automatic extraction of special knowledge graphs designed to address Legal QA through general-purpose deep language models that are not specifically trained on legal documents. In particular, DiscoLQA is composed by the *baseline* tool of Sovrano et al. (2020) extended with a new component responsible for the extraction of special *information units* representing EDUs and AMRs.

The *baseline* tool described by Sovrano et al. (2020) is composed of a pipeline of algorithms for efficient *Question-Answer Retrieval* through the extraction of a knowledge graph from a set of *information units*. In this sense, the main difference between DiscoLQA and the *baseline* is (as shown in Fig. 1) the type of *information units* considered by the knowledge graph extractor. The *baseline* uses as *information units* all the clauses[3] of the source documents.[4] Instead, DiscoLQA can use as *information units* not only the clauses but also the AMRs and discourse relations extracted from the clauses.

In other words, DiscoLQA supports more types of *information units* and allows the retrieval of answers from any combination of clauses, AMRs and discourse

[3] A clause is a group of words that functions as one part of speech and that includes a subject and a verb.
[4] The identification of sentences and clauses in an English text is straightforward with a dependency parser, especially with tools such as Spacy.

relations. Specifically, discourse relations are meant to capture how EDUs are connected, while AMRs are meant to capture the informative components within the EDUs by possibly supporting answering to basic questions such as "who did what to whom, when or where". For example, from the sentence "*The existence and validity of a contract, or any term of a contract, shall be determined by the law which would govern it under this Regulation if the contract or term were valid*" it is possible to extract the following discourse relation about contingency (that we represent as a pair of question and answer for convenience and clarity) "*In what case would the law govern it under this Regulation? If the contract or term were valid*", and the following AMR question-answer "*By what is the existence and validity of a contract determined? The law that would govern it under this Regulation if the contract or clause were valid*". So, a discourse relation identifies two EDUs: the first encoded in the question and the second in the answer.

In this section, we discuss the system implementation of DiscoLQA, starting from the proposed mechanism for extracting EDUs and AMRs.

4.1 Information units extraction: discourse relations and abstract meaning representations

The AMRs and EDUs used by DiscoLQA are extracted from sentences and paragraphs through a deep language model based on T5[5] Raffel et al. (2020) pre-trained on a multi-task mixture of unsupervised and supervised tasks.

Vanilla T5 is not trained to recognise AMRs or EDUs. Therefore we had to fine-tune T5 on some public datasets designed for these tasks. These datasets are namely QAMR (Michael et al. 2018) for extracting AMRs, and QADiscourse (Pyatkin et al. 2020) for EDUs and discourse relations. Interestingly, both datasets encode AMRs and EDUs as question-answer pairs; this is done for convenience only. Indeed, as pointed out by Michael et al. (2018); Pyatkin et al. (2020); Roit et al. (2020) and others, the question-answer format is more natural, facilitating humans to operate changes, correct errors, suggesting improvements, even without knowing in detail all the underlying linguistic theories.

Most importantly, the QAMR and QADiscourse datasets are not related to any of the technical domains covered by Q4EU. They do not contain legal documents or text fragments written in legalese. In other words, by fine-tuning T5 on QAMR and QADiscourse, we do not refine T5 on legal texts. Legal fine-tuning would require the costly extraction of a dataset of AMRs and EDUs from legal texts, also considering *ad hoc* adaptations of discourse theories and abstract meaning representation to legal language.

In particular, the QAMR dataset is made of 107,880 different questions (and answers) that are a mapping of AMR theory to the following wh-phrases:

- What (60.9% of the dataset),

[5] T5 is an encoder-decoder model based on the assumption that all Natural Language Processing problems can be converted in a text-to-text problem.

- Who (17.5%),
- How (6.9%),
- Where (5.0%),
- When (4.3%),
- Which (2.9%),
- Whose (1.9%),
- Why (0.6%).

On the other hand, the QADiscourse dataset is made of 16,613 different questions (and answers) that are a mapping of PDTB to the following wh-phrases mainly on contingency and temporal relations:

- In what manner (25% of the dataset),
- What is the reason (19%),
- What is the result (16%),
- What is an example (11%),
- After what (7%),
- While what (6%),
- In what case (3%),
- Despite what (3%),
- What is contrasted with it (2%),
- Before what (2%),
- Since when (2%),
- What is similar (1%),
- Until when (1%),
- Instead of what (1%),
- What is an alternative ($\leq$ 1%),
- Except when ($\leq$ 1%),
- Unless what ($\leq$ 1%).

The two considered datasets are tuples of $< s, q, a >$, where s is a source sentence, q is a question (implicitly) expressed in s, and a is an answer expressed in s. So that T5 is fine-tuned to tackle at once the following four tasks per dataset:

1. Extract a given s and q,
2. Extract q given s and a,
3. Extract all the possible q given s,
4. Extract all the possible a given s.

Specifically, we fine-tuned the T5 model on QAMR and QADiscourse for five epochs.[6] The objective of the fine-tuning was to minimise a loss function measuring the difference between the expected output (i.e., a for the 1st task, q for the 2nd task,

[6] An epoch is one complete cycle through the entire training dataset.

etc.) and the output given by T5. A mathematical definition of the loss function is given by Raffel et al. (2020).

At the end of the training, the average loss was 0.4098, meaning that our fine-tuned T5 model cannot perfectly extract AMRs or EDUs from the text composing the training set. On the one hand, this is a good thing because it is likely that the model did not over-fit on the training set. On the other hand, this points to the fact that the AMRs and EDUs extracted by our T5 model can be imperfect, containing errors that could propagate to the answer retrieval system. Regardless, in the following sections, we show that even if the language models we rely on are imperfect, we can still outperform the *baseline* information retrieval system.

4.2 System implementation: knowledge graph extraction and answer retrieval

DiscoLQA, similarly to the *baseline* tool described by Sovrano et al. (2020), consists in a pipeline of AI algorithms that is capable of extracting from a set of *information units* a particular graph of knowledge that an information retrieval system can exploit to answer a given question. In particular, this knowledge graph is extracted by detecting, with a dependency parser, all the possible phrases and sub-phrases within the *information units* so that each phrase stands for an edge of the knowledge graph. In practice, these phrases are represented as special triplets of subjects, templates and objects called template-triplets. Specifically, the templates are composed of the ordered sequence of tokens connecting a subject and an object. The subject and the object are represented in such templates with the placeholders *"{subj}"* and *"{obj}"*.

Hence, the resulting template-triplets are a sort of function, where the predicate is the body and the object and the subject are the parameters. Obtaining a natural language representation of these template-triplets is straightforward by design by replacing the instances of the parameters in the body. This natural representation is then used as a possible answer for retrieval by measuring the similarity between its embedding and the embedding of a question. An example of template-triple is:

- Subject: *"the applicable law"*
- Template: *"Surprisingly {subj} is considered to be clearly more related to {obj} rather than to something else"*
- Object: *"that Member State"*

Because of the adopted extraction procedure, the resulting knowledge graph could be better. It may contain mistakes caused by wrongly identified grammatical dependencies or other issues.

To increase the interoperability of the extracted knowledge graph with external resources, we formatted it as an RDF graph. RDF is a standard model for data interchange on the Web (Allemang and Hendler 2011). In particular, RDF has features that facilitate data merging even if the underlying schemas differ. To format a graph of template triplets in an RDF graph, we performed the following steps:

- We assigned a Uniform Resource Identifier (URI) to every node (i.e., subject and object) and edge (i.e., template) of the graph by lemmatising the associated text. To each URI, we assigned an RDFS label corresponding to the associated text.
- We added special triplets to keep track of the sources from which the template-triplets were extracted so that for each node and edge is possible to go back to the source document or paragraph.
- We added sub-class relations between composite concepts (syntagms) and the simplest concepts (if any) composing the syntagm. For example, "*contractual obligation*" is a sub-class of "*obligation*".

For more technical details about how we performed all the steps mentioned above to convert the template-triplets into an RDF graph, please refer to Sovrano et al. (2020) or the source code of DiscoLQA.

Finally, the algorithm to retrieve answers from the extracted knowledge graph is based on the following steps. Let C be the set of concepts in a question q, and $m =< s, t, o >$ be a template-triplet, and $u = t(s, o)$ be the natural language representation of m also called *information unit*, and z its source paragraph. DiscoLQA performs answer retrieval by finding the most similar concepts to C within the knowledge graph, retrieving all their related template-triplets m (including those of the sub-classes), and selecting amongst the natural language representations u of the retrieved template-triplets those that are likely to be an answer to q. The probability that u pertinently answers q can be estimated through SyntagmTuner (Sovrano et al. 2022) as the numerical similarity between the embedding of $u + z$ (i.e., u concatenated with z) and the embedding of q. So that if $u + z$ is similar enough to q, then z is said to be an answer to q for the *information unit* u. Therefore, the algorithm can retrieve any arbitrary number of answers, given that enough information units are available.

In particular, the embeddings of $u + z$ and q are obtained through a deep language model specialised on QA retrieval and pre-trained on ordinary English to associate similar vectorial representations to a question and its correct answers. The pre-trained deep language models we considered for our implementation of DiscoLQA and our experiments are the Universal Sentence Encoder (Yang et al. 2020), MiniLM (Wang et al. 2021), and MPNet (Song et al. 2020).

5 Experiment

Given all the premises stated in Sect. 1 and Sect. 3, we designed an experiment to better understand the role of discourse relations in legalese, in order to determine how to exploit existing state-of-the-art general-purpose natural language models for QA in order to automatically and effectively answer questions on legal documents (e.g., Private International Law). Indeed, legalese is a technical language in many ways similar to its related natural language, but with important differences in how the meaning is encoded in the text. Legalese is not repetitive. It is canonical and has semantic terminology that tends to avoid polysemy and to be used punctually in particular contexts as if the sentences it forms were governed by very formal rules.

We hypothesise that applying these formal rules affects the syntagmatic relationships within sentences and discourse structure. Suppose this hypothesis were correct, in principle, it would be possible to specialise general-purpose natural language models to legalese simply by integrating them with external information about the structure of discourse of legal texts without costly training procedures otherwise hampered by the scarcity of data. This is why we decided to design an experiment focused on understanding whether there is a benefit in using discourse relations and AMRs instead of plain sentences when performing Question-Answer Retrieval on the body[7] of articles and recitals. The overall idea is that using discourse relations and AMRs as *information units* would help to partly crystallise into the retrieval system the structure of discourse used by the legal texts. This would make it invariant, avoiding the answer retriever using the discourse schemes learned from the common language instead.

Hence we designed DiscoLQA that, as described in Sect. 4, extends the *baseline* Question-Answer Retrieval system proposed by Sovrano et al. (2020), supporting different combinations of *information units*, i.e., AMR and discourse relations. So, for the experiment, we can compare the performance of different *information units* on the same answer retrieval algorithm. More precisely, we want to study the following instances of DiscoLQA:

- **Clause**: equivalent to the QA tool by Sovrano et al. (2020). This is DiscoLQA which uses only clauses as *information units*.
- **Clause+EDU+AMR**: DiscoLQA which uses clauses, discourse relations and AMRs as *information units*, all together.
- **Clause+EDU**: DiscoLQA using clauses and discourse relations but not AMRs.
- **Clause+AMR**: DiscoLQA using clauses and AMRs.
- **EDU+AMR**: discourse relations and AMRs.
- **EDU**: discourse relations.
- **AMR**.

As a result, if one type/combination of *information units* would perform better than the others, the gain in performance would be imputed to the only difference between the tools: the type/combination of adopted *information units*. Therefore, if DiscoLQA were better than the *baseline* (Sovrano et al. 2020), we would have some evidence to support our initial hypothesis by measuring the effects of discourse structure on the performance of information retrievers trained on general-purpose natural language.

We consider as a baseline only the answer retrieval system by Sovrano et al. (2020) mainly for two reasons:

[7] Unlike the body, titles/headings (e.g. of articles, sections, chapters) are usually concise (i.e., few words), so we expect DiscoLQA to have minimal impact on a title, because there would be little noise to remove and no discourse relation to capture.

Table 1 Statistics on Q4EU: the column "Art./Rec." counts the number of recitals and articles. The column "Questions" counts the number of different questions, and the column "Tokens per Art./Rec." counts the mean number of tokens per article/recital, and so on. Please note that Q4EU is the sum of Q4PIL, Q4EAW, Q4GDPR and Q4eIDAS

	Questions	Expected answers	Answers per question	Norms	Art./Rec	Tokens	Tokens per Art./Rec
Q4PIL	17: 5 low; 7 normal; 5 high	65	3.82	3	269	27,280	101.41
+ Q4EAW	21: 7 low; 7 normal; 7 high	68	3.23	1	50	8426	168.52
+ Q4GDPR	17: 4 low; 7 normal; 6 high	55	3.23	1	272	45,138	165.94
+ Q4eIDAS	17: 5 low; 7 normal; 5 high	37	2.17	1	129	17,283	133.97
= Q4EU	72: 21 low; 28 normal; 23 high	225	3.12	6	720	98,127	136.28

1. It is the only system we know that can perform legal question-answering without any ad-hoc fine-tuning or training procedure. We do not have an extensive enough dataset to train an end-to-end QA system on specific European legislation; our focus is on zero-shot legal QA (as defined in Sect. 1).
2. It is the only legal question-answer retrieval system we know that has been tested on European legislation. Therefore it is the most suitable baseline for us.

To show that the results generalise across different deep language models, we decided to run the experiments on different state-of-the-art deep neural networks for answer retrieval:

- The *Universal Sentence Encoder Q &A model* (USE, for short), by TensorFlow (Yang et al. 2020, Google);
- *MiniLM* (Wang et al. 2021, Microsoft);
- *MPNet* (Song et al. 2020, Microsoft).

In particular, the last two models were fine-tuned on 215 million question-answer pairs[8] by SBERT (Reimers and Gurevych 2019).

We decided to consider only the models mentioned above because: *i)* they are some of the best general-purpose models for the task on TensorFlow and SBERT (two state-of-the-art repositories for deep neural networks easily accessible through user-friendly APIs); *ii)* deep neural networks for answer retrieval (i.e., models for generating vectorial representations of questions and answers) are different from and less common than models for question answering or answer extraction.

[8] See https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-dot-v1.

Unfortunately, we do not know of any general-purpose open-source deep language model trained specifically on legal answer retrieval. The only exception could be the work by Vold and Conrad (2021), though their language model was trained on privacy policies, and they are usually written in more plain English than European legislation (Table 1).

Finally, in order to evaluate DiscoLQA and perform the experiment, we need a dataset of at least 50[9] relevant questions on European legislation, with known expected answers. Considering that Q4PIL (Sovrano et al. 2021) comprises only 17 questions on Private International Law, we decided to build a larger test set called Q4EU, to include more questions on different European norms, as described in Sect. 5.

6 Q4EU: a test set for legal answer retrieval

Q4EU contains 72 unique questions and 225 expected answers (i.e., articles and recitals). For simplicity of exposition, **Q4EU can be divided into the following sub-sets**:

- **Q4PIL** (see Table 2): containing questions about 3 private international laws: Rome I Regulation EC 593/2008; Rome II Regulation EC 864/2007; Brussels I bis Regulation EU 1215/2012. These regulations are, respectively, on the law applicable to contractual obligations, on the law applicable to non-contractual obligations, on jurisdiction and the recognition and enforcement of judgements in civil and commercial matters. In particular, they aim to provide a tool for identifying the applicable law and the jurisdiction in cases when two or more legal systems connect and generate complex relationships (e.g., a sale of goods contract between an Italian and a German citizen regarding commodities situated in Spain).
- **Q4EAW** (see Table 3): containing questions about the Council Framework Decision (CFD) of 13 June 2002 on the European arrest warrant and the surrender procedures between Member States.[10] In particular, this framework decision increases the efficiency of extradition procedures for crime suspects. Furthermore, it also determines the abolition of formal extradition procedures between member states of the EU for persons who are fugitives from justice after being finally convicted. The framework decision represents the first concretisation of the principle of free movement of judicial decisions in criminal matters, encompassing both pre-sentence and final decisions by fostering judicial cooperation and the development of a single area of freedom, security and justice in the EU.

[9] The minimum number of queries required for a valid information retrieval test set in order to obtain statistically significant results is normally 50 (Clough and Sanderson 2013).

[10] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02002F0584-20090328&from=EN

Table 2 Q4PIL subset: here, "B" stands for Brussels I bis Regulation EU 1215/2012, "RI" for Rome I Regulation EC 593/2008 and "RII" for Rome II Regulation EC 864/2007

Question	Specificity	Expected answers	Target
Who determines disputes under a contract?	L	Art.7.1, Art.8.3, Art.8.4, Art.17	B
What factors should be taken into account for conferring the jurisdiction to determine disputes under a contract?	N	Art.7.1, Art.17, Art.20, Art.25	B
Which parties of a contract should be protected by conflict-of-law rules?	N	Rec.23, Art.6, Art.8, Art.13	RI
In which case are claims so closely connected that it would be better to treat them together in order to avoid irreconcilable judgments?	H	Art.8, Art.30, Art.34	B
What kind of agreement between parties is regulated by these Regulations?	L	B Rec.6, B Rec.10, B Rec.12, B Art.1, RI Rec.7, RI Art.1	B, RI, RII
In which court is celebrated the trial in case the employer is domiciled in a Member State?	H	Art.21, Art.22, Art.23	B
How should a contract be interpreted according to Regulation Rome I?	L	Rec.22, Rec.12, Rec.26, Rec.29, Art.12	RI
Which law is applicable to a non-contractual obligation?	N	Rec.17, Rec.18, Rec.26, Rec.27, Rec.31, Art.4-20	RII
Can the parties choose the applicable law in consumer contracts?	H	Rec.11, Rec.25, Rec.27, Art.6	RI
What factors should be taken into account for conferring the jurisdiction to determine disputes under a consumer contract?	N	Rec.18, Art.17, Art.18, Art.19, Art.26	B
Can the parties choose a different applicable law for different parts of the contract?	L	Rec.11, Art.3.1	RI
What non-contractual obligations fall into the scope of Regulation Rome II?	H	Rec.10, Rec.11, Art.1, Art.2	RII
What is the applicable rule to protect the weaker party of a contract?	N	RI Rec.23, B Rec.18	B, RI
What is the applicable law to determine the validity of consent?	L	Art.3.5, Art.10, Art.11, Art.13	RI
When are two actions to be considered related according to the Regulation Brussels I Bis?	N	Rec.21, Art.30.3	B
What court has jurisdiction in case of a counter-claim?	N	Art.8.3, Art.14.2, Art.18.3, Art.22.2	B
Where can an employee sue their employer?	H	Rec.14, Rec.18, Art.21.1, Art.22.1, Art.23	B

Furthermore, "Rec." stands for Recital, and "Art." for Article. In the "Specificity" column: "L" stands for Low, "N" stands for Normal and "H" stands for High. The "Target" column indicates which norm a question is targeting. If no norm is indicated next to the article/recital, the norm of the article/recital is indicated in the "Target" column

Table 3 Q4EAW subset: here, "W" stands for the CFD of 13 June 2002 on the European arrest warrant and the surrender procedures between Member States

Question	Specificity	Expected answers	Target
What is the European arrest warrant?	N	Art.1.1, Art.8, Art.9.3, Rec.11, Rec.6	W
Can the execution of the European arrest warrant be refused when the law of the executing Member State does not impose the same type of tax or duty or does not contain the same type of tax rules as the law of the issuing Member State?	L	Art.2.2, Art.2.4, Art.4.1, Rec.6	W
Who decides precedence in the event of a conflict between a European arrest warrant and a request for extradition from a third country?	N	Art.16.3, Rec.8, Art.10.6	W
Which law is used to record the consent to surrender of a requested person?	H	Art.13.3, Art.11	W
Is the arrest warrant based on the principle of mutual recognition?	L	Rec.2, Rec.6, Rec.5, Art.1.1, Art.1.2, Rec.10	W
Does a requested person have the right to an interpreter?	H	Art.11.2	W
Can the consent to the surrender of the arrested person be revoked?	N	Art.13.4, Art.17	W
Is the surrender of the arrested person always subject to the verification of the double criminality of the act?	L	Art.2.2, Art.2.3, Art.2.4, Art.4.1, Art.5, Art.33	W
Which authority should be informed in case of repeated delays by a Member State in executing European arrest warrants?	H	Art.17.7	W
Can the Member States also apply other agreements in addition to the Framework Decision?	L	Art.31, Rec.5, Art.33, Art.32	W
Can the European arrest warrant be ordered for the execution of a non-custodial sentence?	N	Art.2.1, Art.1.1, Rec.12, Art.5	W
Can the executing judicial authority refuse to execute the European arrest warrant when the person who is the subject of the European arrest warrant is being prosecuted in the executing Member State for the same act as that on which the European arrest warrant is based?	N	Art.4.2, Rec.8, Art.24, Rec.13	W
What right is applied by the judicial authority to decide whether the requested person should remain in detention or be provisionally released?	H	Art.12.1, Rec.8, Rec.10	W
Can the constitutional rules of the Member States be applied?	L	Rec.7, Rec.12, Art.1.3, Art.34	W
Should the European arrest warrant be translated into the official language or one of the official languages of the executing Member State?	H	Art.8.2, Rec.8	W
Can the executing judicial authority request the opinion of Eurojust in case of multiple requests?	H	Art.16.2, Rec.8	W

Table 3 (continued)

Question	Specificity	Expected answers	Target
Can the executing judicial authority, on its own initiative, seize and hand over property acquired by the requested person as a result of an offence?	N	Art.29.1, Rec.5	W
Is an alert in the Schengen Information System equivalent to a European arrest warrant?	N	Art.9.3, Art.8.1, Art.1.1	W
What are the time limits for the surrender of the requested person?	L	Art.23, Art.15, Art.17, Art.20, Art.24, Rec.1	W
How are the expenses of executing the European arrest warrant allocated?	H	Art.30	W
What claims can be made to the judicial authority by the interested party who has not previously received any official information on the existence of the criminal proceedings against him/her?	L	Art.4a, Rec.12, Art.11	W

For more details on how to read this table, please see the caption of Table 2

Table 4 **Q4GDPR subset**: here, "G" stands for GDPR

Question	Specificity	Expected answers	Target
Does the GDPR provide a right to explanation?	L	Rec.71, Art.12.3, Art.15.1	G
When is it mandatory to carry out a Data Protection Impact Assessment?	H	Art.35.1, Art.35.3	G
What are the possible security measures that can be adopted to mitigate the risks related to personal data processing?	N	Art.32.1, Art.32.2	G
What are the applicable rules to the processing of personal data for archiving purposes in the public interest, for scientific or historical research purposes or for statistical purposes?	L	Rec.156, Art.5.1.b, Art.9.2.j, Art.14.5.b, Art.17.3.d, Art.89	G
How should a data processor be appointed?	N	Art.26, Art.38	G
When is the consent of the data subject explicit?	L	Rec.51, Rec.71, Rec.111, Art.7.1, Art.9	G
What elements shall the European Commission keep into account to authorise the transfer of personal data to a third country through an Adequacy Decision?	N	Art.45.2, Art.45.3, Rec.104	G
What are the rules applicable to biometric data?	H	Rec.51, Rec.53, Art.9	G
When does the public interest override data subject rights?	L	Rec.45, Rec.46, Rec.50, Rec.65, Rec.69, Art.9.2.i, Art.17.3, Art.89	G
To what data is the right to portability applicable?	H	Art.20	G
How should a data processing record be drafted?	H	Art.30	G
What data processing poses significant risks to the fundamental rights and freedoms of natural persons?	N	Rec.51, Rec.75, Art.9, Art.10	G
What elements should be included in a Code of Conduct?	N	Rec.81, Art.40	G
What are the obligations of the data controller when the legal basis for the data processing is the consent of the data subject?	N	Art.7, Art.13, Art.14, Art.20	G
Which legal entity can impose fines on data controllers?	H	Rec.130, Art.58.2.i, Art.83	G
Who can exercise the right to lodge a complaint before the supervisory authority?	N	Rec.141, Rec.142, Art.77	G
What is the procedure to follow in the event of a data breach?	L	Rec.85, Rec.86, Rec.87, Rec.88, Art.33, Art.34	G

For more details on how to read this table, please see the caption of Table 2

Table 5 **Q4eIDAS subset**: here, "E" stands for the eIDAS Regulation

Question	Specificity	Expected answers	Target
How is a qualified electronic signature validated?	H	Art.32, Art.33, Rec.57	E
Can an electronic signature be expressed in the form of a pseudonym?	N	Art.3.14, Art.32	E
Can a minor obtain a qualified electronic signature?	L	Art.3, Art.25	E
From when qualified certificates lose their validity in the case of revocation?	N	Art.24, Art.28	E
Is a graphometric signature qualified as an advanced electronic signature?	L	Art.3.11, Art.26	E
How should access to trust services be granted to persons with disabilities?	N	Rec.29, Art.15	E
How can the identity of a natural person be verified in the issuing of a qualified certificate?	H	Art.24.1	E
Do electronic contracts have the same validity as paper contracts?	L	Rec.21, Art.2.3	E
Why is there a specific discipline for the notification of security breaches?	H	Rec.38, Art.19.2	E
When shall a trust service provider notify affected individuals and users?	H	Art.19.2	E
What is the applicable law to the trust service provider which provides its trusted services in a Member State different from the one where it is established?	L	Rec.22, Rec.42, Art.4, Art.6, Art.24	E
How can qualified certificates be temporally limited?	N	Rec.53, Art.24.4, Art.28, Art.38.5	E
What are the requirements for website authentication?	N	Rec.67, Art.45	E
When do electronic signatures qualify as "advanced electronic signatures"?	N	Art.3.11, Art.26	E
Which subject has the competence to maintain trusted lists?	H	Art.22	E
How should liability be determined for Member States that are non-compliant with provisions about electronic identification schemes?	N	Rec.18, Art.11	E
What is a security breach?	L	Art.10, Art.19	E

For more details on how to read this table, please see the caption of Table 2

- **Q4GDPR** (see Table 4): containing questions about the General Data Protection Regulation (GDPR),[11] the most relevant piece of legislation in the EU legal framework with regards to data protection law. Its goal is to foster the fundamental right to data protection, enshrined by the Charter of Fundamental Rights of the European Union (art. 8), while harmonising rules in data processing, profiling, and risk management.
- **Q4eIDAS** (see Table 5): containing questions about Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC,[12] also known as eIDAS Regulation. This legislation tackles several issues in electronic identification, electronic signature, electronic seals, and trust services. Its goal is to provide legal certainty for cross-border transactions in the EU Single Market.

Some statistics on the datasets mentioned above are shown in Table 1.

To build the Q4EU dataset and, in the first place, the Q4PIL dataset, the pieces of legislation (i.e., the norms) kept into account are conceived as self-contained legal environments. While legal interpretation is often grounded on external legal factors (e.g., jurisprudence, scholars' opinions), we opted for a "black letter" approach to the law that only considers the legislative legal formant. Therefore, the point of view assumed in our analysis is the perspective of the lawmakers. This has a twofold implication for question-and-answer drafting.

On the one hand, questions have been modelled to be answered solely within the legal text under scrutiny. They do not refer to legal concepts, such as the hierarchy of legal sources or competence, that are not explicitly mentioned in the regulations. Moreover, not all the (legal) questions are the same. While some accept as an answer a provision that exactly matches the question, others rely on more complex interpretations (i.e., legal reasoning) to be answered. Therefore, questions have been classified depending on their context specificity, which can either be low, normal, or high.

First, specific questions whose answer is precisely in the domain of the regulations and an answer is provided in the "black letter" of the law were labelled as highly specific. An example of a question with high specificity is "In what court can an employee sue its employer?" because it perfectly falls within the scope and goals of Regulation Brussels I-bis and finds its exact answer in the provisions of Articles 21 and 23.

Questions whose answer falls within the scope of the regulations while requiring an abstraction of multiple legal provisions were labelled as normally specific. For instance, "What is the applicable rule to protect the weaker party of a contract?" was labelled as normally specific since the answer also relies on the concept of "weaker

[11] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons concerning the processing of personal data and the free movement of such data, and repealing Directive 95/46/EC, https://eur-lex.europa.eu/eli/reg/2016/679/oj

[12] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0910

party" mentioned across two regulations (Recital 23 Rome I and Recital 18 Brussels I) concerning any contract (as a legal concept) rather than specific contractual types.

Finally, broad questions whose tentative answer is found through an articulate combination of articles and recitals were labelled as having low specificity. For instance, a question with low specificity is "Can the parties choose a different applicable law for different parts of the contract?". While Rome I Regulation provides for a discipline on the applicable law to contract, it does not contain any provision concerning individual parts. The answer is ultimately open to interpretation in such a question, whereas the Regulation suggests norms that could serve as a reference point.

Since such classification might be subjective and dependent on each jurist, three legal experts independently evaluated the level of context specificity and decided by the majority about the final level.

On the other hand, the answers to the questions provided by legal experts, which constitute the dataset used to observe the performance of deep language modes, are obtained by mirroring the question-drafting methodology. Three legal experts, different from the question-drafters, provided answers to the legal questions by looking for the following:

1. Specific, punctual, and explicit answers in the case of highly specific questions;
2. General and conceptual, yet text-based, answers to normally specific questions; and
3. Prima facie textual references to be used as interpretative points of reference in the case of low specific questions.

These experts only provided textual references in the legislation at the article or recital level (e.g., Rome I art. 8; B Rec. 18). When at least two experts agree on a given answer, their response is valid without further enquiry. If one expert provides another answer, another expert validates this response. In drafting the validation answers, no other articles or recitals have been considered except those provided by the original validators.

7 Results and error analysis

Considering that, with the Q4EU dataset, a single answer is not sufficient[13] to respond to a test query altogether, we relied on top-k precision, F1, Normalised Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR) as evaluation metrics. In particular, the top-k precision, or $P@k$, is measured as the fraction of expected answers amongst the top-k retrieved instances. The top-k F1 score, or $F1@k$, is given by $2\frac{R@k \cdot P@k}{R@k+P@k}$, where the top-k recall, or $R@k$, is measured as the

[13] DiscoLQA and the baseline have no constraints on the minimum or maximum number of retrievable responses.

Table 6 **Q4EU**—scores of universal sentence encoder

Universal Sentence Encoder	Top5 Scores		Top10 Scores	
	All Norms Search	Target Norms Search	All Norms Search	Target Norms Search
Clause (*Baseline*)	P: 0.516 ± 0.323 F1: 0.383 ± 0.202 NDCG: 0.425 ± 0.25 MRR: 0.675 ± 0.373	P: 0.531 ± 0.313 F1: 0.409 ± 0.203 NDCG: 0.47 ± 0.261 MRR: 0.728 ± 0.357	P: 0.614 ± 0.31 F1: 0.331 ± 0.166 NDCG: 0.412 ± 0.226 MRR: 0.686 ± 0.355	P: 0.653 ± 0.294 F1: 0.368 ± 0.174 NDCG: 0.462 ± 0.246 MRR: 0.735 ± 0.342
AMR	P: 0.434 ± 0.309 F1: 0.322 ± 0.209 NDCG: 0.397 ± 0.314 MRR: 0.582 ± 0.417	P: 0.466 ± 0.324 F1: 0.354 ± 0.229 NDCG: 0.433 ± 0.325 MRR: 0.612 ± 0.419	P: 0.537 ± 0.339 F1: 0.29 ± 0.186 NDCG: 0.417 ± 0.314 MRR: 0.587 ± 0.411	P: 0.579 ± 0.342 F1: 0.332 ± 0.203 NDCG: 0.46 ± 0.32 MRR: 0.62 ± 0.409
EDU	P: 0.518 ± 0.325 F1: 0.364 ± 0.205 NDCG: 0.412 ± 0.268 MRR: 0.677 ± 0.366	P: 0.544 ± 0.312 F1: 0.403 ± 0.212 NDCG: 0.467 ± 0.264 MRR: 0.747 ± 0.345	P: 0.635 ± 0.296 F1: 0.353 ± 0.191 NDCG: 0.434 ± 0.245 MRR: 0.686 ± 0.349	P: 0.683 ± 0.283 F1: 0.393 ± 0.196 NDCG: 0.486 ± 0.243 RR: 0.752 ± 0.334
EDU+AMR	P: 0.528 ± 0.32 F1: 0.383 ± 0.204 **NDCG: 0.442 ± 0.279** MRR: 0.694 ± 0.371	P: 0.554 ± 0.314 F1: 0.411 ± 0.211 **NDCG: 0.491 ± 0.287** **MRR: 0.759 ± 0.356**	P: 0.649 ± 0.299 F1: 0.359 ± 0.186 **NDCG: 0.439 ± 0.244** MRR: 0.703 ± 0.356	P: 0.7 ± 0.291 F1: 0.399 ± 0.198 **NDCG: 0.49 ± 0.258** **MRR: 0.765 ± 0.344**
Clause+AMR	**P: 0.547 ± 0.324** **F1: 0.396 ± 0.203** NDCG: 0.423 ± 0.25 MRR: 0.682 ± 0.358	P: 0.557 ± 0.322 F1: 0.412 ± 0.211 NDCG: 0.464 ± 0.268 MRR: 0.74 ± 0.355	P: 0.63 ± 0.312 F1: 0.352 ± 0.19 NDCG: 0.421 ± 0.234 MRR: 0.685 ± 0.353	P: 0.673 ± 0.285 F1: 0.388 ± 0.187 NDCG: 0.469 ± 0.247 MRR: 0.744 ± 0.346
Clause+EDU	P: 0.535 ± 0.339 **F1: 0.396 ± 0.226** NDCG: 0.426 ± 0.26 MRR: 0.688 ± 0.38	**P: 0.562 ± 0.325** **F1: 0.428 ± 0.224** NDCG: 0.47 ± 0.266 MRR: 0.741 ± 0.352	**P: 0.666 ± 0.301** **F1: 0.368 ± 0.187** NDCG: 0.417 ± 0.222 MRR: 0.699 ± 0.361	**P: 0.701 ± 0.291** **F1: 0.402 ± 0.195** NDCG: 0.463 ± 0.237 MRR: 0.746 ± 0.342
Clause+EDU+AMR	P: 0.526 ± 0.323 F1: 0.39 ± 0.215 NDCG: 0.425 ± 0.258 **MRR: 0.697 ± 0.37**	P: 0.541 ± 0.32 F1: 0.411 ± 0.22 NDCG: 0.468 ± 0.275 MRR: 0.751 ± 0.36	P: 0.665 ± 0.293 **F1: 0.368 ± 0.185** NDCG: 0.419 ± 0.221 **MRR: 0.705 ± 0.357**	P: 0.696 ± 0.287 F1: 0.399 ± 0.194 NDCG: 0.467 ± 0.24 MRR: 0.759 ± 0.344

This table shows the macro mean (with standard deviation) of the top-k precision (P), F1, NDCG and MRR of each combination of information units. We show the values for $k = \{5, 10\}$, either considering all the norms (when retrieving the answers) or considering only the documents for which the questions were designed. The best column scores are shown in bold, while darker background colours indicate higher precision column-wise

fraction of correct answers retrieved in the top-k instances. In contrast, the top-k NDCG (Sakai 2007) is a measure of ranking quality normalised in [0, 1] that measures the usefulness, or gain, of an answer based on its position in the result list. Instead, the top-k MRR (Voorhees 1999) only cares about the single highest-ranked relevant item. It shows what system does the best job at placing a relevant document/passage in to highest rank.

It is important to note that the main difference between precision, F1, MRR and NDCG is that the last two are used to assess the ability of an answer retrieval system to rank correct answers first. Conversely, the other metrics measure the system's precision and accuracy. For these reasons, all selected metrics are considered complementary measurements that may present different lenses into the problem of understanding answer retrieval systems (Dato et al. 2022).

In Tables 6, 7 and 8 we show the macro[14] top-k evaluation scores for $k = \{5, 10\}$,[15] studying how different types of *information units* and *deep language models*

[14] Here, the term "macro" means that precision, F1, NDCG and MRR scores are computed independently for each test query and then averaged, to put an equal weight upon the contribution of each query.

[15] In general, a k greater than or equal to the average number of answers per question (e.g., the score shown in Table 1) is recommended.

Table 7 **Q4EU**—scores of MiniLM

MiniLM	Top5 Scores		Top10 Scores	
	All Norms Search	Target Norms Search	All Norms Search	Target Norms Search
Clause (*Baseline*)	P: 0.539 ± 0.318 F1: 0.415 ± 0.207 **NDCG: 0.478 ± 0.257** MRR: 0.741 ± 0.344	P: 0.549 ± 0.314 F1: 0.431 ± 0.218 NDCG: 0.512 ± 0.268 MRR: 0.786 ± 0.332	P: 0.643 ± 0.3 F1: 0.372 ± 0.187 NDCG: 0.453 ± 0.225 MRR: 0.746 ± 0.335	P: 0.669 ± 0.294 F1: 0.402 ± 0.193 NDCG: 0.499 ± 0.242 MRR: 0.79 ± 0.322
AMR	P: 0.456 ± 0.324 F1: 0.34 ± 0.222 NDCG: 0.423 ± 0.318 MRR: 0.625 ± 0.418	P: 0.488 ± 0.328 F1: 0.374 ± 0.227 NDCG: 0.472 ± 0.329 MRR: 0.673 ± 0.414	P: 0.56 ± 0.333 F1: 0.308 ± 0.192 NDCG: 0.447 ± 0.305 MRR: 0.631 ± 0.409	P: 0.564 ± 0.331 F1: 0.329 ± 0.204 NDCG: 0.485 ± 0.324 MRR: 0.675 ± 0.41
EDU	P: 0.492 ± 0.321 F1: 0.365 ± 0.214 NDCG: 0.423 ± 0.281 MRR: 0.667 ± 0.385	P: 0.528 ± 0.322 F1: 0.41 ± 0.231 NDCG: 0.476 ± 0.291 MRR: 0.707 ± 0.368	P: 0.628 ± 0.308 F1: 0.348 ± 0.19 NDCG: 0.439 ± 0.256 MRR: 0.677 ± 0.369	P: 0.663 ± 0.305 F1: 0.377 ± 0.193 NDCG: 0.485 ± 0.259 MRR: 0.713 ± 0.358
EDU+AMR	P: 0.527 ± 0.319 F1: 0.394 ± 0.212 NDCG: 0.454 ± 0.277 MRR: 0.705 ± 0.37	P: 0.554 ± 0.306 F1: 0.432 ± 0.221 NDCG: 0.507 ± 0.286 MRR: 0.747 ± 0.355	**P: 0.688 ± 0.292** F1: 0.38 ± 0.185 **NDCG: 0.459 ± 0.243** MRR: 0.713 ± 0.355	P: 0.726 ± 0.279 F1: 0.413 ± 0.187 **NDCG: 0.506 ± 0.249** MRR: 0.755 ± 0.34
Clause+AMR	P: 0.54 ± 0.317 F1: 0.412 ± 0.204 NDCG: 0.471 ± 0.256 MRR: 0.741 ± 0.334	P: 0.55 ± 0.315 F1: 0.433 ± 0.219 NDCG: 0.518 ± 0.271 **MRR: 0.798 ± 0.322**	P: 0.652 ± 0.3 F1: 0.375 ± 0.189 NDCG: 0.451 ± 0.224 MRR: 0.742 ± 0.33	P: 0.676 ± 0.294 F1: 0.403 ± 0.198 NDCG: 0.498 ± 0.244 MRR: 0.8 ± 0.317
Clause+EDU	**P: 0.562 ± 0.314** **F1: 0.434 ± 0.197** NDCG: 0.471 ± 0.243 MRR: 0.728 ± 0.341	**P: 0.577 ± 0.31** **F1: 0.456 ± 0.209** NDCG: 0.51 ± 0.255 MRR: 0.772 ± 0.331	P: 0.679 ± 0.287 F1: 0.386 ± 0.182 NDCG: 0.449 ± 0.218 MRR: 0.731 ± 0.334	P: 0.726 ± 0.279 **F1: 0.421 ± 0.189** NDCG: 0.492 ± 0.23 MRR: 0.775 ± 0.323
Clause+EDU+AMR	P: 0.549 ± 0.316 F1: 0.421 ± 0.2 NDCG: 0.472 ± 0.244 **MRR: 0.752 ± 0.33**	P: 0.569 ± 0.309 F1: 0.452 ± 0.208 **NDCG: 0.52 ± 0.254** **MRR: 0.798 ± 0.322**	P: 0.683 ± 0.289 **F1: 0.39 ± 0.181** NDCG: 0.457 ± 0.214 **MRR: 0.756 ± 0.322**	**P: 0.727 ± 0.283** **F1: 0.421 ± 0.191** NDCG: 0.499 ± 0.228 **MRR: 0.802 ± 0.312**

For further details on interpreting this table, read the caption of Table 6

Table 8 **Q4EU**—scores of MPNet

MPNet	Top5 Scores		Top10 Scores	
	All Norms Search	Target Norms Search	All Norms Search	Target Norms Search
Clause (*Baseline*)	P: 0.529 ± 0.309 F1: 0.413 ± 0.209 NDCG: 0.466 ± 0.259 MRR: 0.744 ± 0.348	P: 0.566 ± 0.3 F1: 0.449 ± 0.211 **NDCG: 0.507 ± 0.266** MRR: 0.778 ± 0.331	P: 0.666 ± 0.272 F1: 0.375 ± 0.173 NDCG: 0.452 ± 0.221 MRR: 0.751 ± 0.334	P: 0.694 ± 0.265 F1: 0.409 ± 0.189 NDCG: 0.497 ± 0.24 MRR: 0.783 ± 0.319
AMR	P: 0.452 ± 0.33 F1: 0.337 ± 0.227 NDCG: 0.412 ± 0.311 MRR: 0.602 ± 0.405	P: 0.457 ± 0.327 F1: 0.351 ± 0.232 NDCG: 0.445 ± 0.324 MRR: 0.65 ± 0.412	P: 0.553 ± 0.341 F1: 0.299 ± 0.194 NDCG: 0.427 ± 0.295 MRR: 0.608 ± 0.397	P: 0.569 ± 0.34 F1: 0.323 ± 0.203 NDCG: 0.463 ± 0.309 MRR: 0.654 ± 0.407
EDU	P: 0.502 ± 0.319 F1: 0.373 ± 0.213 NDCG: 0.434 ± 0.273 MRR: 0.703 ± 0.376	P: 0.539 ± 0.327 F1: 0.408 ± 0.23 NDCG: 0.476 ± 0.281 MRR: 0.729 ± 0.365	P: 0.641 ± 0.313 F1: 0.346 ± 0.195 NDCG: 0.447 ± 0.255 MRR: 0.715 ± 0.355	P: 0.67 ± 0.292 F1: 0.375 ± 0.197 NDCG: 0.488 ± 0.264 MRR: 0.74 ± 0.343
EDU+AMR	P: 0.522 ± 0.301 F1: 0.391 ± 0.212 NDCG: 0.46 ± 0.286 MRR: 0.732 ± 0.364	P: 0.561 ± 0.303 F1: 0.426 ± 0.22 NDCG: 0.506 ± 0.284 **MRR: 0.782 ± 0.343**	P: 0.667 ± 0.303 F1: 0.37 ± 0.193 **NDCG: 0.458 ± 0.243** MRR: 0.74 ± 0.35	P: 0.693 ± 0.293 F1: 0.393 ± 0.202 NDCG: 0.493 ± 0.256 MRR: 0.785 ± 0.336
Clause+AMR	P: 0.529 ± 0.316 F1: 0.4 ± 0.204 NDCG: 0.463 ± 0.258 **MRR: 0.751 ± 0.342**	P: 0.558 ± 0.316 F1: 0.43 ± 0.215 NDCG: 0.506 ± 0.268 MRR: 0.8 ± 0.329	P: 0.65 ± 0.294 F1: 0.371 ± 0.187 NDCG: 0.452 ± 0.224 **MRR: 0.756 ± 0.331**	P: 0.682 ± 0.28 F1: 0.407 ± 0.191 **NDCG: 0.501 ± 0.24** **MRR: 0.806 ± 0.316**
Clause+EDU	P: 0.546 ± 0.312 F1: 0.414 ± 0.208 NDCG: 0.454 ± 0.25 MRR: 0.741 ± 0.345	P: 0.576 ± 0.307 F1: 0.448 ± 0.214 NDCG: 0.497 ± 0.258 MRR: 0.776 ± 0.333	**P: 0.685 ± 0.282** **F1: 0.389 ± 0.179** NDCG: 0.449 ± 0.21 MRR: 0.748 ± 0.331	**P: 0.715 ± 0.269** **F1: 0.423 ± 0.189** NDCG: 0.492 ± 0.229 MRR: 0.782 ± 0.321
Clause+EDU+AMR	**P: 0.562 ± 0.311** **F1: 0.426 ± 0.202** **NDCG: 0.467 ± 0.254** MRR: 0.74 ± 0.344	**P: 0.581 ± 0.307** **F1: 0.453 ± 0.215** NDCG: 0.506 ± 0.264 MRR: 0.775 ± 0.336	P: 0.667 ± 0.296 F1: 0.387 ± 0.191 NDCG: 0.451 ± 0.219 MRR: 0.743 ± 0.338	P: 0.702 ± 0.282 F1: 0.422 ± 0.198 NDCG: 0.494 ± 0.235 MRR: 0.778 ± 0.328

For further details on interpreting this table, read the caption of Table 6

Table 9 **Q4EU**—average length of information units by type

	Clauses	AMRs	Discourse relations
Mean length	32.39	24.98	30.96

This table shows the average number of characters of the discourse relations, AMRs and clauses used by DiscoLQA and the baseline

Table 10 Statistical tests

P@10 on USE	Precision		F1		NDCG		MRR	
(Targeted search)	*T*	*p* value	*T*	*p* value	*T*	*p* value	*T*	*p* value
EDU+AMR	**130.0**	**0.04**	**294.5**	**0.06**	806.0	0.08	157.0	0.21
EDU	138.5	0.11	456.5	0.08	861.5	0.08	170.0	0.32
AMR	413.5	0.89	1061.5	0.94	1146.0	0.43	622.0	0.98
EDU+AMR+Clause	**66.0**	**0.04**	**149.0**	**0.01**	766.0	0.31	135.0	0.22
EDU+Clause	**17.0**	**0.004**	**62.0**	**0.001**	670.0	0.34	141.0	0.39
AMR+Clause	73.5	0.19	**80.5**	**0.04**	581.5	0.37	98.0	0.39

The table reports the Wilcoxon's statistic *T* and the *p* value for the top10 scores ("target norms search") of each combination of information unit and evaluation metric. Statistically significant results ($p < 0.05$) are highlighted in bold

affect answer retrieval. In particular, we show two different evaluations in these tables. The first one is performed by running the answer retrieval algorithm on all the 6 norms of Q4EU (we will refer to it as "all norms search"), even though the questions in Q4EU usually target only 1 or 2 norms. Instead, the second one (we will refer to it as "target norms search") is performed by considering only the legal acts targeted by every question (e.g., Q4GDPR targets only the GDPR, Q4eIDAS only eIDAS), filtering out all the answers coming from unrelated norms.

As expected, all the scores obtained with a "target norms search" are higher than with a "all norms search". Interestingly, the difference between the two evaluations clearly shows the weight of incorrect selection of the target document with DiscoLQA in Q4EU. Nonetheless, these results show that regardless of the choice of *k*, using discourse relations (EDUs) as *information units* gives the best precision, especially when in combination with clauses and AMRs.

Despite their differences, MPNet, MiniLM (the best) and the Universal Sentence Encoder behave very similarly, suggesting that the information units we considered may play a role independent from the underlying language model used for retrieval. DiscoLQA using only discourse relations and AMRs as information units (i.e., *EDU+AMR*) outperforms the baseline in terms of precision. This happens with all the language models considered, except MPNet. This fact suggests that EDUs and AMRs can retain most of the relevant information of the corpus of technical documents, supporting our hypothesis. Moreover, as shown in Table 9, the average length of EDUs and AMRs is smaller than that of normal clauses, further corroborating the hypothesis and demonstrating that the deep language models considered can be distracted by longer clauses.

Table 11 Statistical tests

P@10 on MiniLM	Precision		F1		NDCG		MRR	
(Targeted search)	T	p value	T	p value	T	p value	T	p value
EDU+AMR	**77.0**	**0.03**	275.5	0.12	947.0	0.50	216.5	0.85
EDU	175.5	0.50	509.0	0.84	1109.0	0.75	308.0	0.97
AMR	375.0	0.99	1109.5	0.99	1176.0	0.67	386.0	0.98
EDU+AMR+Clause	**17.0**	**0.004**	**68.0**	**0.02**	696.0	0.62	72.5	0.28
EDU+Clause	**6.0**	**0.002**	**60.0**	**0.04**	816.0	0.81	68.0	0.67
AMR+Clause	9.0	0.19	43.5	0.44	494.0	0.70	25.5	0.25

See the caption of 10 for more details on how to interpret this table

In light of the similarities and differences observed across different algorithms and information units, a statistical test was essential to ascertain the significance of these findings. Since the data samples considered are not independent, we opted for the Wilcoxon signed-rank test (Woolson 2007), a non-parametric version of the paired T-test that is suitable for paired samples. Indeed, the same questions are tested across all algorithms (i.e., EDU, Clause, etc.).

The results of the one-sided statistical tests on the top10 scores ("target norms search") of the Universal Sentence Encoder and MiniLM are shown[16] respectively in Tables 10 and 11. Statistically significant improvements were generally seen in the precision and F1 scores when using a combination of EDUs, AMRs and clauses. MiniLM showed significant gains mainly in precision, whereas the Universal Sentence Encoder displayed more widespread improvements, particularly in F1 scores. Neither answer retriever exhibited statistically significant changes in NDCG and MRR metrics.

Overall, these findings support our hypothesis. They show that it is possible to improve a general-purpose language model, making it perform better with legal texts. This is possible by better capturing syntagmatic relationships and using noiseless *information units*, i.e., decomposing a generic clause into one or more discourse relations or AMRs.

In other words, as expected, the information units representing the (generic) clauses carry enough noise to distract the answer retriever. By breaking the sentences into EDUs and explicitly keeping their relations, we can crystallise the discourse structure into the knowledge graph, making it invariant. Therefore the answer retriever is forced to "reason" over the discourse patterns, minimising the chances of relying on common-sense discourse schemes instead.

Examples of how EDUs and AMRs are important for some questions of the Q4EU dataset are shown in Table 12. In particular, a qualitative analysis of the algorithm's responses shows that it can identify useful normative references to ensure the completeness of the answer and develop an overview. For example, among the

[16] The results for MPNet are not shown because none of them was significant.

Table 12 **Q4EU**—examples of correct answers

Question	Algorithm	Answer	Answer text	Answer pertinence (%)	Key IU	Type of IU
What court has jurisdiction in case of a counter-claim?	EDU+AMR	B Art. 8.3	A person domiciled in a Member State may also be sued: (3) on a counter-claim arising from the same contract or facts on which the original claim was based, in the court in which the original claim is pending;	49.35	May a person domiciled in a Member State be sued? in the court in which the original claim is pending	Discourse Relation
Where can an employee sue their employer?	EDU+AMR	B Art. 23.2	Which allows the employee to bring proceedings in courts other than those indicated in this Section.	25.81	What allows the employee to bring proceedings? courts	AMR
Who determines disputes under a contract?	EDU+AMR	B Art. 17.2	Where a consumer enters into a contract with a party who is not domiciled in a Member State but has a branch, agency or other establishment in one of the Member States, that party shall, in disputes arising out of the operations of the branch, agency or establishment, be deemed to be domiciled in that Member State.	35.26	Who enters into a contract with a party? a consumer	AMR
What court has jurisdiction in case of a counter-claim?	EDU+AMR	B Art. 14.2	The provisions of this Section shall not affect the right to bring a counter-claim in the court in which, in accordance with this Section, the original claim is pending.	59.30	Iis the right to bring a counter-claim? in the court	Discourse Relation

Table 12 (continued)

Question	Algorithm	Answer	Answer text	Answer pertinence (%)	Key IU	Type of IU
What factors should be taken into account for conferring the jurisdiction to determine disputes under a contract?	Baseline	B Art. 25.5	An agreement conferring jurisdiction which forms part of a contract shall be treated as an agreement independent of the other terms of the contract.	34.41	The terms of the contract	Clause
What kind of agreement between parties are regulated by these Regulations?	Baseline	RI Art. 1.2.e	The following shall be excluded from the scope of this Regulation: (e) arbitration agreements and agreements on the choice of court;	37.18		Clause

This table shows a few examples of answers correctly given by the baseline and DiscoLQA (EDU+AMR). For each answer, we provide the key *information unit* (IU for short) used by the retriever to associate the correct answer to the question. Nonetheless, even if the answer is selected correctly, the reason for selecting it might be wrong. Thus, errors in the *information units* are shown with a strike-through

Table 13 **Q4EU**—examples of wrong answers: this table shows a few examples of answers wrongly given by the baseline and DiscoLQA (EDU+AMR)

Question	Algorithm	Answer	Answer text	Answer pertinence (%)	Key IU	Type of IU
What kind of agreement between parties is regulated by these Regulations?	Baseline	B Art. 73.3	This Regulation shall not affect the application of bilateral conventions and agreements between a third State and a Member State concluded before the date of entry into force of Regulation (EC) No 44/2001 which concern matters governed by this Regulation.	45.16	Of conventions and agreements	Clause
When are two actions to be considered related according to Regulation Brussels I Bis?	EDU+AMR	B Art. 71.2.a	The court hearing the action shall, in any event, apply Article 28 of this Regulation;	27.53	In what case shall the court hearing the action apply Article 28 of the Regulation? In any event	Discourse Relation
Can the parties choose a different applicable law for different parts of the contract?	EDU+AMR	R1 Rec. 14	Should the Community adopt, in an appropriate legal instrument, rules of substantive contract law, including standard terms and conditions, such instrument may provide that the parties may choose to apply those rules.	41.28	What may the parties choose to apply? substantive contract law	AMR

For more details on this table, read the caption of Table 12

Table 14 **Q4EU**—P@10 by context specificity

P@10 on MiniLM (Targeted Search)	Specificity		
	High	Normal	Low
Clause (*Baseline*)	0.781 ± 0.295	0.645 ± 0.31	0.586 ± 0.23
AMR	0.531 ± 0.382	0.612 ± 0.327	0.535 ± 0.267
EDU	0.783 ± 0.28	0.631 ± 0.313	0.584 ± 0.281
EDU+AMR	0.828 ± 0.225	**0.702 ± 0.315**	0.654 ± 0.246
Clause+AMR	0.758 ± 0.297	0.686 ± 0.312	0.58 ± 0.234
Clause+EDU	**0.842 ± 0.222**	0.688 ± 0.302	0.658 ± 0.265
Clause+EDU+AMR	**0.842 ± 0.222**	0.685 ± 0.31	**0.667 ± 0.266**

Mean top10 precision scores (with standard deviation) grouped by *context specificity*, for MiniLM with "target norms search". The best column results are in bold, while darker background colours indicate higher precision column-wise

Table 15 Q4EU—percentage of answers more/less precise than the baseline

P@10 on MiniLM (Targeted Search)	Specificity		
	High	Normal	Low
AMR	More: 9.09% Less: 45.45%	More: 21.43% Less: 17.86%	More: 9.09% Less: 27.27%
EDU	More: 18.18% Less: 13.64%	More: 14.29% Less: 17.86%	More: 22.73% Less: 22.73%
EDU+AMR	**More: 18.18%** **Less: 9.09%**	**More: 25.0%** **Less: 7.14%**	**More: 27.27%** **Less: 9.09%**
Clause+AMR	More: 0.0% Less: 4.55%	**More: 17.86%** **Less: 0.0%**	More: 0.0% Less: 4.55%
Clause+EDU	**More: 9.09%** **Less: 0.0%**	More: 14.29% Less: 0.0%	More: 22.73% Less: 9.09%
Clause+EDU+AMR	**More: 9.09%** **Less: 0.0%**	More: 21.43% Less: 7.14%	**More: 22.73%** **Less: 4.55%**

Percentage of queries for which DiscoLQA (with MiniLM, "target norms search") made a positive/negative difference from the baseline in terms of top10 precision. Percentages are grouped by *context specificity*. The best column *deltas* are in bold, while darker background colours indicate higher positive *deltas* (the difference between "more" and "less") column-wise

answers to the question "Who decides precedence in the event of a conflict between a European arrest warrant and a request for extradition from a third country?" the algorithm identifies Article 16.3 (the most relevant answer) and suggests Recital 8, which helps interpret Article 16.3. Furthermore, for the same question, the algorithm also suggests Article 10.6, which, while not suitable for answering the question, leads the jurist to complementary points of reference for more holistic reasoning and interpretation.

Both Tables 12 and 13 show errors committed by the answer retrievers and the extractor of information units. These examples clearly reveal at least two different types of errors. The first type occurs when an information unit is extracted to be semantically or grammatically incorrect, such as in the first and fourth rows of Table 12. This type of error is relatively minor since, in some cases, the underlying

language model is resistant to inaccuracies[17], still allowing a correct answer to be retrieved, as shown in 12.

In particular, this first type of error is usually caused by the automatic extraction of AMRs and EDUs by a neural network, as described in Sect. 4.1. For this reason, it is possible to see in both Tables 12 and 13 examples of information units that do not perfectly overlap with the text of a response. On the other hand, the second type of error is due to mistakes in the deep language model for answer retrieval. As shown in Table 13, this type of error can be rather severe, causing wrong answers to be selected by the retriever.

As in the evaluation carried out by Sovrano et al. (2021), we studied how (top-10) precision scores vary when the *context specificity* changes. Results partly confirm our *expectations*. We can see a trend where mean top-10 precision increases proportionally to the *context specificity*. This is clear in all instances of DiscoLQA, except AMR. In particular, as shown in Table 14, AMRs only contribute to better answer questions having low and normal specificity. Furthermore, we also show in Table 15 the percentage of queries for which DiscoLQA made a positive/negative difference from the baseline in terms of top-10 precision and grouped by specificity.

Our *expectations* were based on the fact that:

- The specificity of a question is low when it asks something that cannot be explicitly found in the Regulations but requires a holistic analysis of principles, competence rules, and so forth;
- Questions with low specificity usually tend to have more expected answers, and it may be harder to find all of them;
- Multi-hop reasoning is usually required to answer questions with low specificity, but the considered answer retrievers are not equipped for that kind of reasoning (yet).

For example, the question "How should a contract be interpreted according to Regulation Rome I?" has a very low specificity. It requires pinpointing both recitals and articles for a proper answer, therefore, more distinct and distant paragraphs. Most of the questions regarding hermeneutics would probably require a broader view of the subject, having a low specificity to the Regulation, therefore requiring multi-hop reasoning.

8 Discussion and conclusion

With this paper, we empirically investigated the role of discourse structure in legalese, trying to understand its importance in encoding the meaning of legal documents. Ours is a first attempt to exploit more sophisticated linguistic theories such as PDTB. To this end, we devised a simple experiment on legal question answering,

[17] This mainly happens because the underlying language model relies on the contextualised embedding of the information unit (and not just that of the unit alone), as explained at the end of Sect. 4.2.

designed to shed more light on whether Elementary Discourse Units (EDUs) and Abstract Meaning Representations (AMRs) are the fundamental information units in legislative texts as well.

As a result of these experiments, we found that EDUs and AMRs seem to be useful for better capturing long-distant relations between information units, as shown in Table 15. This leads to an overall improvement of our DiscoLQA over the baseline, in terms of precision, F1, NDCG and MRR. In particular, EDU+AMR (the version of DiscoLQA using AMRs and EDUs) was able to produce 23.61% more precise top10 answers than the baseline, using MiniLM with "target norms search". This percentage rises to 25% and 27.27% when considering only questions with normal and low specificity, respectively.

The goal of our experiments was also practical, not just theoretical. Understanding how legalese differs from its natural language can help us address the problem of data scarcity in legalese processing/understanding by allowing us to exploit general-purpose language models not specifically trained on legal documents. However, these generic language models may be one of many available. Indeed, in the literature, it is possible to find several examples of training data for legal domains, or at least training data that can be exploited via transfer learning paradigms. Nonetheless, transfer learning is challenging, and different legal domains or documents may deploy different discourse structures, requiring different language models. For example, privacy policies can be considered legal documents, though their language is usually closer to plain English than legalese, to help consumers understand the policy. In other words, transfer learning can be an alternative solution to zero-shot question answering. However, neither of the two approaches can be considered a one-size-fits-all solution for all possible problems.

We tested and evaluated DiscoLQA on specific European norms and a relatively small dataset without comparing our results with deep language models pre-trained on legal corpora, as explained at the end of Sect. 5. Nonetheless, even though Q4EU is about different legal sub-domains (respectively: Private International Law, the European arrest warrant, data protection and electronic signatures), our instances of DiscoLQA were able to generalise well across them, outperforming the baseline in all the cases. Notably, this result occurred even though we built DiscoLQA to perform zero-shot question answering without any training procedure involving European legislation or (more generally) legal documents. Therefore, DiscoLQA can potentially be used in various domains where data scarcity is unavoidable. To implement DiscoLQA, it is not necessary to manually create a new, time-consuming dataset, such as Q4EU.

Another discussion we should have is about the scalability of DiscoLQA. Indeed, DiscoLQA introduces some extra overhead on the identified baseline, but this overhead does not affect either the asymptotic time complexity of answer retrieval or pre-processing. More precisely, the time complexity of pre-processing changes only by a constant factor. This is because EDUs and AMRs are extracted in polynomial time from paragraphs (and not documents) by a pre-trained deep neural network that does not need to be retrained in order to work. Furthermore, the time complexity of retrieval can only increase by a constant factor, i.e., when EDUs and AMRs are combined with normal clauses. This is because the number and size of EDUs and

AMRs normally never exceed that of clauses. Even when only EDUs or AMRs are considered instead of clauses, the time complexity is reduced by the smaller number of information units to be searched.

In most of today's deep learning applications, the test and training sets are much larger than those used in these experiments. For example, the MS Marco (Nguyen et al. 2016) collection (partly also used for training MiniLM and MPNet) consists of over 1 million questions whose answers are extracted from 3.5 million web documents. These large datasets only make sense for training and evaluating generic language models on tasks that do not suffer from data scarcity. In these cases, due to bandwidth and scalability issues, a pre-processing strategy such as that employed by DiscoLQA and the baseline could introduce a significant memory overhead into the information retrieval system. Instead, due to the small size of the Q4EU dataset (less than 300 items per sub-collection), we can easily implement an extractor of knowledge graphs (and other relationship identifiers).

On the one hand, working with less data poses several technical challenges that sometimes require paradigm shifts. On the other hand, it can also open the way for several technological solutions previously considered impractical. In this article, we have shown only a few examples of how deep learning strategies can be rethought to adapt to smaller data and problems. We have only scratched the tip of an iceberg that may be uncovered by emerging ideas from joint efforts in the field of AI and law. For instance, as future work, we point to the possibility of specialising the algorithm for extracting EDUs and AMRs to legislative texts, taking into account what we already know about legal connectors and discourses.

Acknowledgements F. Sovrano gratefully acknowledge the support of the Swiss National Science Foundation through the SNF Project 200021_197227.

Author Contributions F. Sovrano: conceptualization, methodology, data curation, original draft preparation, investigation, validation, formal analysis, software, visualization. M. Palmirani: supervision, data curation methodology. S. Sapienza: data curation (Q4GDPR and Q4eIDAS datasets). V. Pistone: data curation (Q4EAW dataset) and the related error analysis (see Sect. 7).

Funding Open access funding provided by University of Zurich.

References

Allemang D, Hendler JA (2011) Semantic web for the working ontologist - effective modeling in RDFS and owl, second edition. Morgan Kaufmann. Accessed from http://www.elsevierdirect.com/product.jsp?isbn=9780123859655

Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Schneider N (2013) Abstract meaning representation for sembanking. In: Dipper S, Liakata M, Pareja-Lora A (eds), Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, law-id@acl 2013, August 8–9, 2013, Sofia, Bulgaria, pp 178–186. The Association for Computer Linguistics. Accessed from https://aclanthology.org/W13-2322/

Cabrio E, Tonelli S, Villata S (2013) From discourse analysis to argumentation schemes and back: Relations and differences. In: Leite J, Son TC, Torroni P, van der Torre L, Woltran S (eds), Computational logic in multi-agent systems - 14th international workshop, CLIMA xiv, Corunna, Spain, Sept 16–18, 2013. Proceedings, vol 8143. Springer, pp 1–17. Accessed from https://doi.org/10.1007/978-3-642-40624-9_1

Cao ND, Aziz W, Titov I (2019) Question answering by reasoning across documents with graph convolutional networks. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, Jun 2–7, 2019, vol 1 (long and short papers). Association for Computational Linguistics, pp 2306–2317. Accessed from https://doi.org/10.18653/v1/n19-1240

Chalkidis I, Kampas D (2019) Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 27(2):171–198

Chen D, Zhang S, Zhang X, Yang K (2020) Cross-lingual passage re-ranking with alignment augmented multilingual BERT. IEEE Access 8:213232–213243. https://doi.org/10.1109/ACCESS.2020.3041605

Chen D, Yih W (2020) Open-domain question answering. In: Savary A, Zhang Y (eds), Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts, ACL 2020, online, July 5, 2020. Association for Computational Linguistics, pp 34–37. Accessed from https://doi.org/10.18653/v1/2020.acl-tutorials.8

Clough PD, Sanderson M (2013) Evaluating the performance of information retrieval systems using test collections. Inf Res 18:2

Condevaux C, Harispe S, Mussard S, Zambrano G (2019) Weakly supervised one-shot classification using recurrent neural networks with attention: application to claim acceptance detection. In: Araszkiewicz M, Rodríguez-Doncel V (eds) Legal knowledge and information systems - JURIX 2019: The thirtysecond annual conference, Madrid, Spain, Dec 11–13, 2019, vol 322. IOS Press, pp 23–32. Accessed from https://doi.org/10.3233/FAIA190303

D'Angelo F (1984) Nineteenth-century forms/modes of discourse: a critical inquiry. Coll Comp Commun 35(1):31–42

Dato D, MacAvaney S, Nardini FM, Perego R, Tonellotto N (2022) The istella22 dataset: Bridging traditional and neural learning to rank evaluation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G (eds) SIGIR '22: The 45th international ACM SIGIR conference on research and development in information retrieval, Madrid, Spain, July 11–15, 2022. ACM, pp 3099–3107. Accessed from https://doi.org/10.1145/3477495.3531740

Fludernik M (2000) Genres, text types, or discourse modes? narrative modalities and generic categorization. Style 34(2):274–292

Gordon TF, Walton D (2009) Legal reasoning with argumentation schemes. In: The 12th international conference on artificial intelligence and law, proceedings of the conference, June 8–12, 2009, Barcelona, Spain, pp 137–146. ACM. Accessed from https://doi.org/10.1145/1568234.1568250

Hage J (2000) Defeasible deontic logic. Artif Intell Law 8(1):75–91

Hudson DA, Manning CD (2019) GQA: a new dataset for real-world visual reasoning and compositional question answering. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, Long beach, CA, USA, June 16–20, 2019. Computer Vision Foundation/IEEE, pp 6700–6709. https://doi.org/10.1109/CVPR.2019.00686

ICLR (2019) Blackstone. Accessed 09 Mar 2021 from https://research.iclr.co.uk. (**Online**)

Bommarito II MJ, Katz DM, Detterman EM. LexNLP: Natural language processing and information extraction for legal and regulatory texts. InResearch handbook on big data law 2021 May 14 (pp. 216-227). Edward Elgar Publishing. http://arxiv.org/abs/1806.03688

Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Yih W-t (2020) Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). Online: Association for Computational Linguistics, pp 6769–6781. Accessed from https://aclanthology.org/2020.emnlp-main.550

Kim M, Xu Y, Goebel R (2015) Applying a convolutional neural network to legal question answering. In: Otake M, Kurahashi S, Ota Y, Satoh K, Bekki D (eds), New frontiers in artificial intelligence

- jsai-isai 2015 workshops, lenls, jurisin, aaa, hat-mash, tsdaa, asd-hr, and skl, Kanagawa, Japan, Nov 16–18, 2015, revised selected papers, vol 10091, pp 282–294. Accessed from https://doi.org/10.1007/978-3-319-50953-2_20

Lam H, Governatori G (2009) The making of spindle. In: Governatori G, Hall J, Paschke A (eds), Rule interchange and applications, international symposium, ruleml 2009, Las Vegas, Nevada, USA, November 5–7, 2009. Proceedings, vol 5858. Springer, pp 315–322. Accessed from https://doi.org/10.1007/978-3-642-04985-9_29

Lascarides A, Asher N (2007) Segmented discourse representation theory: dynamic semantics with discourse structure. In: Bunt H, Muskens R (eds), Computing meaning. Springer Netherlands, Dordrecht, pp 87–124. Accessed from https://doi.org/10.1007/978-1-4020-5958-2_5

Mann WC, Thompson SA (1988) Rhetorical structure theory: toward a functional theory of text organization. Text Interdiscipl J Study Disc 8(3):243–281

Michael J, Stanovsky G, He L, Dagan I, Zettlemoyer L (2018) Crowdsourcing question-answer meaning representations. In: Walker MA, Ji H, Stent A (eds), Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, Naacl-Hlt, New Orleans, Louisiana, USA, June 1–6, 2018, vol 2 (short papers). Association for Computational Linguistics, pp 560–568. Accessed from https://doi.org/10.18653/v1/n18-2089

Miltsakaki E, Prasad R, Joshi AK, Webber BL (2004) The penn discourse treebank. In: Proceedings of the fourth international conference on language resources and evaluation, LREC 2004, May 26–28, 2004, Lisbon, Portugal. European Language Resources Association. Accessed from http://www.lrecconf.org/proceedings/lrec2004/summaries/618.htm

Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: a human generated machine reading comprehension dataset. In: Besold TR, Bordes A, d'Avila Garcez AS, Wayne G (eds) Proceedings of the workshop on cognitive computation: integrating neural and symbolic approaches 2016 co-located with the 30th annual conference on neural information processing systems (NIPS 2016), Barcelona, Spain, Dec 9 2016, vol 1773. CEUR-WS.org. Accessed from http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

Prasad R, Bunt H (2015) Semantic relations in discourse: the current state of ISO 24617-8. In: Proceedings of the 11th joint ACL-ISO workshop on interoperable semantic annotation (ISA-11). Association for Computational Linguistics, London, UK. Accessed from https://aclanthology.org/W15-0210

Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi AK, Webber BL (2008) The penn discourse treebank 2.0. In: Proceedings of the international conference on language resources and evaluation, LREC 2008, 26 May–1 June 2008, Marrakech, Morocco. European Language Resources Association. Accessed from http://www.lrecconf.org/proceedings/lrec2008/summaries/754.html

Pyatkin V, Klein A, Tsarfaty R, Dagan I (2020) Qadiscourse—discourse relations as QA pairs: Representation, crowdsourcing and baselines. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, No 16–20, 2020. Association for Computational Linguistics, pp 2804–2819. Accessed from https://doi.org/10.18653/v1/2020.emnlp-main.224

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-totext transformer. J Mach Learn Res 21:1401–14067

Ravichander A, Black AW, Wilson S, Norton TB, Sadeh NM (2019) Question answering for privacy policies: combining computational and legal perspectives. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, Nov 3–7, 2019. Association for Computational Linguistics, pp 4946–4957. Accessed from https://doi.org/10.18653/v1/D19-1500

Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, Nov 3–7, 2019. Association for Computational Linguistics, pp 3980–3990. Accessed from https://doi.org/10.18653/v1/D19-1410

Robaldo L, Villata S, Wyner A, Grabmair M (2019) Introduction for artificial intelligence and law: special issue "natural language processing for legal texts''. Artif Intell Law 27(2):113–115

Robaldo L, Miltsakaki E, Hobbs JR (2008) Refining the meaning of sense labels in PDTB: "concession". In: Bos J, Delmonte R (eds) Semantics in Text Processing. STEP 2008 Conference Proceedings, Venice, Italy, Sept 22–24, 2008. Association for Computational Linguistics. Accessed from https://aclanthology.org/W08-2217/

Roit P, Klein A, Stepanov D, Mamou J, Michael J, Stanovsky G, Dagan I (2020) Controlled crowdsourcing for high-quality QA-SRL annotation. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, Jul 5–10, 2020. Association for Computational Linguistics, pp. 7008–7013. Accessed from https://doi.org/10.18653/v1/2020.acl-main.626

Roy U, Constant N, Al-Rfou R, Barua A, Phillips A, Yang Y (2020) LAReQA: Language-agnostic answer retrieval from a multilingual pool. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (emnlp). Online Association for Computational Linguistics, pp 5919–5930. Accessed from https://aclanthology.org/2020.emnlp-main.477

Sakai T (2007) On the reliability of information retrieval metrics based on graded relevance. Inf Process Manag 43(2):531–548

Sanders TJ, Spooren WP, Noordman LG (1992) Toward a taxonomy of coherence relations. Discourse Process. 15(1):1–35

Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, Ma S (2020) BERTPLI: modeling paragraph-level interactions for legal case retrieval. In: Bessiere C (ed) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp 3501–3507. ijcai.org. Accessed from https://doi.org/10.24963/ijcai.2020/484

Song K, Tan X, Qin T, Lu J, Liu T (2020) Mpnet: masked and permuted pre-training for language understanding. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, neurips 2020, Dec 6–12, 2020, virtual. Accessed from https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html

Sovrano F, Palmirani M, Vitali F. Combining shallow and deep learning approaches against data scarcity in legal domains. Government Information Quarterly. 2022 Jul 1;39(3):101715.

Sovrano F, Palmirani M, Distefano B, Sapienza S, Vitali F (2021) A dataset for evaluating legal question answering on private international law. In: Maranhão J, Wyner AZ (eds), ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21–25, 2021. ACM, pp 230–234. Accessed from https://doi.org/10.1145/3462757.3466094

Sovrano F, Palmirani M, Vitali F (2020) Legal knowledge extraction for knowledge graph based question-answering. In: Villata S, Harasta J, Kremen P (eds) Legal Knowledge and Information Systems—JURIX 2020: The Thirty-Third Annual Conference, Brno, Czech Republic, Dec 9–11, 2020, vol 334. IOS Press, pp 143–153. Accessed from https://doi.org/10.3233/FAIA200858

Stede M (2013) Discourse processing. In: Vanderwende L, III HD, Kirchhoff K (eds), Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9–14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. The Association for Computational Linguistics, pp 4–6. Accessed from https://aclanthology.org/N13-4002/

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc, Red Hook, NY, USA, pp 6000–6010. Accessed from https://doi.org/10.5555/3295222.3295349

Vink M, Netten N, Bargh MS, van den Braak SW, Choenni S (2020) Mapping crime descriptions to law articles using deep learning. In: Charalabidis Y, Cunha MA, Sarantis D (eds) ICEGOV 2020: 13th International Conference on Theory and Practice of Electronic Governance, Athens, Greece, 23–25 Sept 2020. pp 33–43. ACM. Accessed from https://doi.org/10.1145/3428502.3428507

Vold A, Conrad JG (2021) Using transformers to improve answer retrieval for legal questions. In: Maranhão J, Wyner AZ (eds) ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, Jun 21–25, 2021. ACM, pp 245–249. Accessed from https://doi.org/10.1145/3462757.3466102

Voorhees EM (1999) The TREC-8 question answering track report. In: Voorhees EM, Harman DK (eds), Proceedings of the Eighth Text Retrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, Nov 17–19, 1999, vol 500–246. National Institute of Standards and Technology (NIST). Accessed from http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf

Wang W, Bao H, Huang S, Dong L, Wei F (2021) Minilmv2: multi-head self-attention relation distillation for compressing pretrained transformers. In: Zong C, Xia F, Li W, Navigli R (eds) Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, online event, Aug 1–6, 2021, vol ACL/IJCNLP 2021. Association for Computational Linguistics, pp 2140–2151. Accessed from https://doi.org/10.18653/v1/2021.findings-acl.188

Webber B, Prasad R, Lee A, Joshi A (2019) The Penn Discourse Treebank 3.0 Annotation Manual. University of Pennsylvania, Philadelphia

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Rush AM (2020) Transformers: State-of-the-art natural language processing. In: Liu Q, Schlangen D (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020—demos, online, Nov 16–20, 2020. Association for Computational Linguistics, pp 38–45. Accessed from https://doi.org/10.18653/v1/2020.emnlpdemos.6

Woolson RF (2007) Wilcoxon Signed-Rank Test. Wiley Encyclopedia of Clinical Trials, 1–3

Xie Z, Thiem S, Martin J, Wainwright E, Marmorstein S, Jansen PA (2020) Worldtree V2: a corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In: Calzolari N et al. (eds) Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11–16, 2020. European Language Resources Association, pp 5456–5473. Accessed from https://aclanthology.org/2020.lrec-1.671/

Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, Kurzweil R (2020) Multilingual universal sentence encoder for semantic retrieval. In: Celikyilmaz A, Wen T (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, online, Jul 5–10, 2020. Association for Computational Linguistics, pp. 87–94. Accessed from https://doi.org/10.18653/v1/2020.acl-demos.12

Zhang Y, Dai H, Kozareva Z, Smola AJ, Song L (2018) Variational reasoning for question answering with knowledge graph. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (aaai-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018. AAAI Press, pp 6069–6076. Accessed from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16983

Zheng L, Guha N, Anderson BR, Henderson P, Ho DE (2021) When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings. In: Maranhão J, Wyner AZ (eds), ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, Jun 21–25, 2021. ACM, pp 159–168. Accessed from https://doi.org/10.1145/3462757.3466088

Zufferey S, Degand L (2017) Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguist Linguist Theory 13(2):399–422

Authors and Affiliations

Francesco Sovrano[1,2] · Monica Palmirani[3] · Salvatore Sapienza[3] · Vittoria Pistone[3]

✉ Francesco Sovrano
francesco.sovrano@uzh.ch

Monica Palmirani
monica.palmirani@unibo.it

Salvatore Sapienza
salvatore.sapienza@unibo.it

Vittoria Pistone
vittoria.pistone@unibo.it

[1] Department of Informatics, University of Zurich, Zurich, Switzerland

[2] DISI, Università di Bologna, Bologna, Italy

[3] CIRSFID-AI, Università di Bologna, Bologna, Italy