Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines

Łukasz Borchmann and **Dawid Wiśniewski** and **Andrzej Gretkowski**
Izabela Kosmala and **Dawid Jurkiewicz** and **Łukasz Szałkiewicz**
Gabriela Pałka and **Karol Kaczmarek** and **Agnieszka Kaliska** and **Filip Graliński**

Applica.ai, Warsaw, Poland
{firstname.surname}@applica.ai

Abstract

We propose a new shared task of semantic retrieval from legal texts, in which a so-called *contract discovery* is to be performed–where legal clauses are extracted from documents, given a few examples of similar clauses from other legal acts. The task differs substantially from conventional NLI and shared tasks on legal information extraction (e.g., one has to identify text span instead of a single document, page, or paragraph). The specification of the proposed task is followed by an evaluation of multiple solutions within the unified framework proposed for this branch of methods. It is shown that state-of-the-art pretrained encoders fail to provide satisfactory results on the task proposed. In contrast, Language Model-based solutions perform better, especially when unsupervised fine-tuning is applied. Besides the ablation studies, we addressed questions regarding detection accuracy for relevant text fragments depending on the number of examples available. In addition to the dataset and reference results, LMs specialized in the legal domain were made publicly available.

1 Introduction

Processing of legal contracts requires significant human resources due to the complexity of documents, the expertise required and the consequences at stake. Therefore, a lot of effort has been made to automate such tasks in order to limit processing costs–notice that law was one of the first areas where electronic information retrieval systems were adopted (Maxwell and Schafer, 2008).

Enterprise solutions referred to as *contract discovery* deal with tasks, such as ensuring the inclusion of relevant clauses or their retrieval for further analysis (e.g., risk assessment). Such processes can consist of a manual definition of a few examples, followed by conventional information retrieval. This approach was taken recently by Nagpal et al. (2018) for the extraction of fairness policies spread across agreements and administrative regulations.

Task	Legal	SI	Few-shot
COLIEE	+	−	−
SNLI	−	−	−
MultiNLI	−	−	−
TREC Legal Track	+	−	−
Propaganda detection	−	+	−
THUMOS (video)	−	+	+
ActivityNet (video)	−	+	+
ALBAYZIN (audio)	−	+	−
Contract Discovery (ours)	+	+	+

Table 1: Comparison of existing shared tasks. Most of the related NLP tasks do not assume Span Identification (*SI*), even those outside the legal domain (*Legal*). Moreover, the few-shot setting is not popular within the field of NLP yet.

2 Review of Existing Datasets

Table 1 summarizes main differences between available challenges. It is shown that most of the related NLP tasks do not assume span identification, even those outside the legal domain. Moreover, the few-shot setting is not popular within the field of NLP yet.

None of existing tasks involving semantic similarity methods, such as SNLI (Bowman et al., 2015) or multi-genre NLI (Bowman et al., 2015), assume span identification. Instead, standalone sentences are provided to determine their entailment. It is also the case of existing shared tasks for legal information extraction, such as COLIEE (Kano et al., 2017), where one has to recognize entailment between articles and queries, as considered in the question answering problem. Obviously, the tasks aimed at retrieving documents consisting of multiple sentences, such as TREC legal track (Baron

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4254–4268

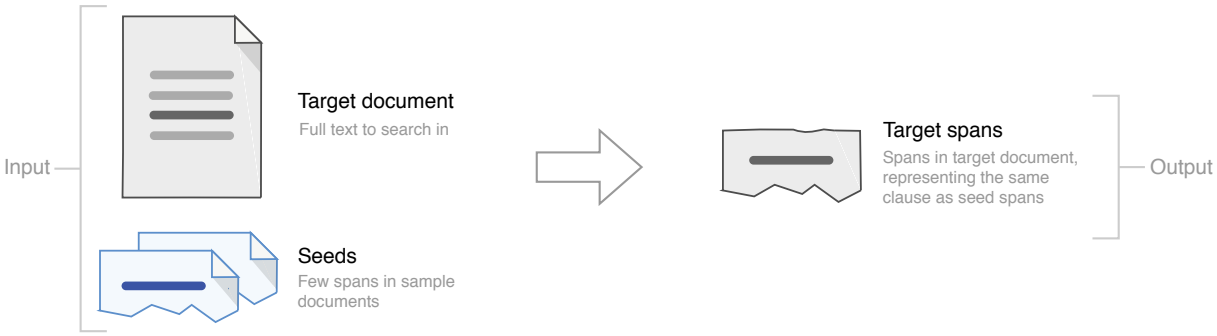

Figure 1: The aim of this task is to identify spans in the requested documents (referred to as *target* documents) representing clauses analogous to the spans selected in other documents (referred to as *seed* documents).

et al., 2006; Oard et al., 2010; Chu, 2011), lack this component.

There are a few NLP tasks where span identification is performed. These include some of plagiarism detection competitions (Potthast et al., 2010) and recently introduced SemEval task of propaganda techniques detection (Da San Martino et al., 2020). When different media are considered, NLP span identification task is equivalent to the action recognition in temporally untrimmed videos where one is expected to provide the start and end times for detected activity. These include THUMOS 14 (Jiang et al., 2014) as well as ActivityNet 1.2 and ActivityNet 1.3 challenges (Fabian Caba Heilbron and Niebles, 2015). Another example is query-by-example spoken term detection, as considered e.g., in ALBAYZIN 2018 challenge (Tejedor et al., 2019).

In a typical business case of *contract discovery* one may expect only a minimal number of examples. The number of available annotations results from the fact that *contract discovery* is performed constantly for different clauses, and it is practically impossible to prepare data in a number required by a conventional classifier every time. When one is interested in the few-shot setting, especially querying by multiple examples, there are no similar shared tasks within the field of NLP. Some authors however experimented recently with few-shot Named Entity Recognition (Fritzler et al., 2019) or few-shot text classification (Bao et al., 2019). The first, however, involves identification of short spans (from one to few words), whereas the second does not assume span identification at all.

What is important, existing tasks aimed at recognizing textual entailment in natural language (Bowman et al., 2015), differ in terms of the domain. This also applies to a multi-genre NLI (Williams et al., 2017), since legal texts vary significantly from other genres. As it will be shown later, methods optimal for MultiNLI do not perform well on the proposed task.

3 Contract Discovery: New Dataset and Shared Task

In this section, we introduce a new dataset of *Contract Discovery*, as well as a derived few-shot semantic retrieval shared task.

3.1 Desiderata

We define our desiderata as follows. We wish to construct a dataset for testing the mechanisms that detect various types of regulations in legal documents. Such systems should be able to process unstructured text; that is, no legal documents segmentation into the hierarchy of distinct (sub)sections is to be given in advance. In other words, we want to provide natural language streams lacking formal structure, as in most of the real-word usage scenarios (Vanderbeck et al., 2011). What is more, it is assumed that a searched passage can be any part of the document and not necessarily a complete paragraph, subparagraph, or a clause. Instead, the process should be considered as a span identification task.

We intend to develop a dataset for identifying spans in a query-by-example scenario instead of the setting where articles are being returned as an answer for the question specified in natural language.

We wish to propose using this dataset in a few-shot scenarios, where one queries the system using multiple examples rather than a single one. The

intended form of the challenge following these requirements is presented in Figure 1. Roughly speaking, the task is to identify spans in the requested documents (referred to as *target* documents) representing clauses analogous (i.e. semantically and functionally equivalent) to the examples provided in other documents (referred to as *seed* documents).

3.2 Data Collection and Annotation

Random subsets of bond issue prospectuses and non-disclosure agreement documents from the US EDGAR database[1], as well as annual reports of charitable organizations from the UK Charity Register[2] were annotated. Note there are no copyright issues and both datasets belong to the public domain.

Annotation was performed in such a way that clauses of the same type were selected (e.g., determining the governing law, merger restrictions, tax changes call, or reserves policy). Clause types depend on the type of a legal act and can consist of a single sentence, multiple sentences or sentence fragments. The exact type of a clause is not important during the evaluation since no full-featured training is allowed and a set of only a few sample clauses can be used during execution.

We restricted ourselves to 21 types as a result of a trade-off between annotation cost and the ability to formulate general remarks. Note that each clause type must be well-understood by the annotator (we described each very carefully in the instructions), and one must have all of the considered clauses in mind when the legal acts are being read during the process. In real-world legal applications, the clauses change in an everyday manner and depend on the problem analyzed by the layer at the moment.

Each document was annotated by two experts, and then reviewed (or resolved) by a super-annotator, who also decided the gold standard. An average Soft F_1 score (Section 4.2) of the two primary annotators, when compared to the gold standard (after the super-annotation), was taken to estimate human baseline performance of 0.84.

The inter-annotator agreement was equal to 0.76 in terms of Soft F_1 metric (Section 4.2). It should be treated as an agreement between two randomly picked annotations since the total number of annotators was 10 (annotators were aligned randomly to a subset of documents in such a way that there would be two annotations and super-annotation per document).

Table 3 presents examples of clauses annotated in the sub-group of Charity Annual Reports documents. The detailed list of clauses and their examples can be found in Appendix C.

The dataset is made publicly available. In addition, we release a large, cleaned, plain-text corpus of legal and financial texts for the purposes of unsupervised model training or fine-tuning. All the available documents of US EDGAR as for November 19, 2018 were crawled. The resulting corpus consists of approx. 1M documents and 2B words in total (1.5G of text after `xz` compression).

3.3 Core Statistics

More than 2,500 spans were annotated in around 600 documents representing either bond issue prospectuses, non-disclosure agreement documents or annual reports of charitable organizations (the detailed statistics regarding the dataset are presented in Table 2).

Annotated clauses differ substantially from what can be found in existing sentence entailment challenges in terms of sentence length and complexity. SNLI contains less than 1% of sentences longer than 20 words, MultiNLI 5%, whereas in the case of clauses, we expect to return and consider it is 93% (and 77% of all spans in our shared task are longer than 20 words).

3.4 Evaluation Framework

Documents were split into halves to form validation and test sets for the purposes of few-shot semantic retrieval challenge. Evaluation is performed by means of a repeated random sub-sampling validation procedure. Sub-samples (k-combinations for each of 21 clauses, $k \in [2, 6]$) drawn from a particular set of annotations are split into $k - 1$ *seed* documents and 1 *target* document. Thus, clauses similar to the *seed* are expected to be returned from the target. We observed that the choice of input examples have an immense impact on the score. It is thus far more important to evaluate various *seed* configurations that various target documents. On the other hand, we wanted to keep the computational cost of evaluation reasonably small, so either the number of seed configurations had to be

[1] http://www.www.sec.gov/edgar.shtml

[2] http://www.gov.uk/find-charity-information

Statistic	
Documents annotated	586
Mean document length (words)	24,284
Clause types	21
Mean clause length (words)	110
Clause instances	2,663

Table 2: Core statistics regarding released dataset.

reduced or the number of target documents for each configuration.

The selected k interval results in 1-shot to 5-shot learning, considered to be few-shot learning (Wang et al., 2019), whereas with the chosen number of sub-samples we expect improvements of 0.01 F_1 to be significant. Note that the 1–5 range denotes the number of annotated documents available, and it is possible that the same clause type appeared twice in one document, resulting in a higher number of clause instances.

Soft F_1 metric on character-level spans is used for the purpose of evaluation, as implemented in *GEval* tool (Graliński et al., 2019). Roughly speaking, this is the conventional F_1 measure, with precision and recall definitions altered to reflect the partial success of returning entities. In the case of the expected clause ranging between $[1, 4]$ characters and the answer with ranges $[1, 3]$, $[10, 15]$ (the system assumes a clause occurs twice within the document), recall equals 0.75 (since this is the part of the relevant item selected) and precision equals ca. 0.33 (since this is the number of selected characters which turned out to be relevant). The Hungarian algorithm (Burkard et al., 2012) is employed to solve the problem of expected and returned range assignments. Soft F_1 has the desired property of being based on the widely utilized F_1 metric while abandoning the binary nature of the match, which is undesirable in the case dealt with in the task described.

4 Competitive Baselines

Solutions based on networks consuming pairs of sequences, such as BERT in sentence pair classification task setting (Devlin et al., 2018a), are considered out of the scope of this paper since they are suboptimal in terms of performance–they require expensive encoding of all combinations from the Cartesian product between seeds and targets, making such solutions unsuitable for semantic similarity search due to the combinatorial explosion (Reimers and Gurevych, 2019). Because of the aforementioned problem and the fact that conventional classifiers require much more data than available in a few-shot setting, in this section, we describe simple k-NN-based approaches that we propose as baseline solutions to the problem stated.

4.1 Processing Pipeline

Evaluated solutions assume pre-encoding of all candidate segments and can be described within the unified framework consisting of segmenters, vectorizers, projectors, aggregators, scorers, and choosers ordered in a pipeline of transformations.

Segmenter is used to split a text into candidate sub-sequences to be encoded and considered in further steps. All the described solutions rely on a candidate sentence and n-grams of sentences, determined with the *spaCy* CNN model trained on OntoNotes.[3] *Vectorizer* produces vector representations of texts on either word, sub-word, or segment (e.g., sentence) level. In our case, vectorization was based on TF-IDF representations, static word embeddings, and neural sentence encoders. *Projector* projects embeddings into a different space (e.g., decomposition methods such as PCA or ICA). *Aggregator* has the capability to use word or sub-word unit embeddings to create a segment embedding (e.g., embedding mean, inverse frequency weighting, autoencoder). *Scorer* compares two or more embeddings and returns computed similarities. Since we often compare multiple seed embeddings with one embedding of a candidate segment, a scorer includes policies to aggregate scores obtained for multiple seeds into the final candidate score (e.g., mean of individual cosine similarities or max-pooling over Word Mover Distances). *Chooser* determines whether to return a candidate segment with a given score (e.g., threshold, one best per document, or a combination thereof). For the sake of simplicity, during the evaluation, we restricted ourselves to the chooser returning only one, the most similar candidate. It is not optimal (because multiple might be expected), but we consider this setting a good reference for further methods.

The proposed taxonomy is consistent with the assumptions made by Gillick et al. (2018). It is presented in order to highlight the similarities and differences between particular solutions when they are introduced and compared within the ablation

[3] http://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-2.1.0

Clause (Instances)	Example
MAIN OBJECTIVE (195/231) The main objective of a charitable organization.	The aim of the Scout Association is to promote the development of young people in achieving their full physical, intellectual, social and spiritual potentials, as individuals, as responsible citizens and as members of their local, national and international communities. The method of achieving the Aim of the Association is by providing an enjoyable and attractive scheme of progressive training based on the Scout Promise and Law and guided by Adult leadership.
GOVERNING DOCUMENT (160/174) Information about the legal document which represents the rule book for the way in which a charity operates (title, date of creation etc.).	The Open University Students Educational Trust (OUSET) is controlled by its governing document, a deed of trust, dated 22 May 1982 as amended by a scheme dated 9 October 1992 and constitutes an unincorporated charity.
TRUSTEE APPOINTMENT (153/168) Procedures for selecting trustees and the term of office.	As per the governing document, four of the Trustee positions are appointed by virtue of their position within the Open University Students Association (OUSA). One further position is appointed by virtue of their previous position within OUSA. One Trustee is nominated by the Vice Chancellor of the Open University (OU) and there are co-opted positions whereby the Trustees are empowered to approach up to two other persons to act as Trustees. It is envisaged that all Trustees will serve a general term of two years in line with the main election periods within OUSA.
RESERVES POLICY (170/185) What are the current financial reserves of the organization and how much these reserves should be as assumed?	The Trustees regularly reviews the amount of reserves that are required to ensure that they are adequate to fulfill the charities continuing obligations.
INCOME SUMMARY (124/134) General information on income for the last year, sometimes associated with information on expenses.	Excluding the adjustments for FRS17 in respect of Pension Fund the results by way of net incoming resources accumulated f3.85m as against E6.78m in 2014, however last years performance benefited from extraordinary property sales generating a profit of F3.15m.
AUDITOR OPINION (190/192) Summary of the opinion of an independent auditor or inspector, often in the form of a list of points.	In connection with my examination, no matter has come to my attention: 1. which gives me reasonable cause to believe that in any material respect the requirements to keep accounting records in accordance with Section 130 of the Charities Act; and to prepare accounts which accord with the accounting records and comply with the accounting requirements of the Charities Act have not been met; or 2. to which, in my opinion, attention should be drawn in order to enable a proper understanding of the accounts to be reached.

Table 3: Clauses annotated in Charity Annual Reports (one of three groups of documents included in the shared task). The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively. More examples are available in Appendix C.

studies later in this paper. The next section describes vectorizers, aggregators, and scorers used for evaluation.

4.1.1 Vectorizers

We intend to provide results of TF-IDF representations, as well as two methods that may be considered the state of the art of sentence embedding. The latter include *Universal Sentence Encoder* (USE) and *Sentence-BERT*.

USE is a Transformer-based encoder, where an element-wise sum of word representations is treated as a sentence embedding (Cer et al., 2018), trained with the multi-task objective. *Sentence-BERT* is a modification of the pretrained BERT network, utilizing Siamese and triplet network structures to derive sentence embeddings, trained with the explicit objective of making them comparable with cosine similarity (Reimers and Gurevych, 2019). In both cases the original models released by the authors were used for the purposes of evaluation.

In addition, multiple contextual embeddings from Transformer-based language models, as well as static (context-less) GloVe word embeddings were tested (Pennington et al., 2014). Many approaches to generating context-dependent vector representations have been proposed in recent years (e.g., Peters et al. (2018); Vaswani et al. (2017)). One important advantage over static embeddings is the fact that every occurrence of the same word is assigned a different embedding vector based on the context in which the word is used. Thus, it is much easier to address issues arising from pre-

trained static embeddings (e.g., taking into consideration polysemy of words). For the purposes of evaluation, we relied on Transformer-based models provided by authors of particular architectures, utilizing the Transformers library (Wolf et al., 2019). These include BERT (Devlin et al., 2018b), GPT-1 (Radford, 2018), GPT-2 (Radford et al., 2018), and RoBERTa (Liu et al., 2019). They differ substantially and introduce many innovations, though they are all based on either the encoder or the decoder from the original model proposed for sequence-to-sequence problems (Vaswani et al., 2017). Selected models were fine-tuned on using the next word prediction task on the Edgar corpus we release and re-evaluated.

4.1.2 Aggregators

In addition to conceptually simple methods such as average or max-polling operations, multiple solutions to utilizing word embeddings for comparing documents can be used. In addition to embeddings mean we evaluated the *Smooth Inverse Frequency* (SIF), *Word Mover's Distance* (WMD) and *Discrete Cosine Transform* (DCT).

SIF is a method proposed by Arora et al. (2017), where a representation of a document is obtained in two steps. First, each word embedding is weighted by $a/(a + f_r)$, where f_r stands for the underlying word's relative frequency, and a is the weight parameter. Then, the projections on the first tSVD-calculated principal component are subtracted, providing final representations.

WMD is a method of calculating a similarity between documents. For two documents, embeddings calculated for each word (e.g., with GloVe) are matched between documents, so that semantically similar pairs of words between documents are detected. This matching procedure generally leads to better results than simply averaging over embeddings for documents and calculating similarity between centers of mass of documents as their similarity (Kusner et al., 2015). Recently, Zhao et al. (2019) showed it might be beneficial to use the method with contextual word embeddings.

DCT is a way to generate document-level representations in an order-preserving manner, adapted from image compression to NLP by Almarwani et al. (2019). After mapping an input sequence of real numbers to the coefficients of orthogonal cosine basis functions, low-order coefficients can be used as document embeddings, outperforming vector averaging on most tasks, as shown by the authors.

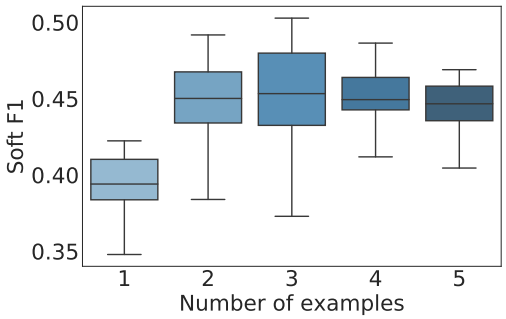

Figure 2: Performance as a function of the number of example documents available (solutions based on LMs). The methods benefit substantially from availability of a second example document and a bigger number leads to a decreased variance.

4.2 Results

Table 4 recapitulates the most important results of the completed evaluation.

Sentence-BERT and Universal Sentence Encoder could not outperform the simple TF-IDF approach, especially when SVD decomposition was applied (the setting commonly referred to as Latent Semantic Analysis). Static word embeddings with SIF weighting performed similarly to TF-IDF, or better, provided they were trained on a legal text corpus rather than on general English. It could not be clearly confirmed whether the use of WMD or DCT is beneficial. For the latter, the best results were achieved with c^0, which in the case of the k-NN algorithm leads to the same answers as mean-pooling and thus is not reported in the table. In case of $c^{0:n}$ where $n > 0$ constant decrease of k-NN methods performance was observed (Appendix B).

Interestingly, from all the released USE models, the multilingual ones performed best — for the monolingual *universal-sentence-encoder-large* model, scores were ten percentage points lower. The best Sentence-BERT model performed significantly worse than the best USE—note that the authors of Sentence-BERT compared it to monolingual models released earlier, which they indeed outperform. Moreover, Sentence-BERT does not perform better than BERT trained with whole word masking, although there is no Sentence-BERT equivalent of this model available so far.

[4]TF-IDF with truncated SVD decomposition is commonly referred to as Latent Semantic Analysis (Halko et al., 2011).

[5]SVD in SIF method is used to perform removal of single common component (Arora et al., 2017).

Segmenter	Vectorizer	Projector	Scorer	Aggregator	Soft F_1
sentence	TF-IDF (1–2 grams, binary TF term)	—	mean cosine	—	0.38
		tSVD (500)[4]	mean cosine	—	**0.39**
sentence	GloVe (300d, Wikipedia & Gigaword)	—	mean cosine	mean	0.34
		—	mean WMD	—	0.35
		SIF tSVD[5]	mean cosine	SIF	0.37
sentence	GloVe (300d, EDGAR)	—	mean cosine	mean	0.36
		—	mean WMD	—	0.35
		SIF tSVD	mean cosine	SIF	**0.41**
sentence	Sentence-BERT (base-nli-stsb-mean ⋆)	—	mean cosine	mean	0.32
sentence	USE (multilingual ⋆)	—	mean cosine	—	**0.38**
sentence	BERT, last layer (large-uncased-whole…⋆)	—	mean cosine	mean	0.35
sentence	GPT-1, last layer	—	mean cosine	mean	0.36
sentence	GPT-2, last layer (large ⋆)	—	mean cosine	mean	0.41
sentence	RoBERTa, last layer (large ⋆)	—	mean cosine	mean	0.31
sentence	GPT-1, last layer (fine-tuned)	—	mean cosine	mean	0.43
sentence	GPT-1, last layer (fine-tuned)	fICA (500)	mean cosine	mean	0.44
sentence	GPT-2, last layer (large, fine-tuned)	—	mean cosine	mean	0.44
sentence	GPT-2, last layer (large, fine-tuned)	fICA (400)	mean cosine	mean	0.45
1–3 sen.	GPT-1, last layer (fine-tuned)		mean cosine	mean	0.47
1–3 sen.	GPT-1, last layer (fine-tuned)	fICA (500)	mean cosine	mean	0.49
1–3 sen.	GPT-2, last layer (large, fine-tuned)		mean cosine	mean	0.46
1–3 sen.	GPT-2, last layer (large, fine-tuned)	fICA (400)	mean cosine	mean	**0.51**
human					**0.84**

Table 4: Selected results when returning a single, most similar segment, determined with given segmenters, vectorizers, projectors, scorers and aggregators. The ⋆ symbol indicates only the best models from each architecture are presented here (results for the remaining ones are available in Appendix B).

In cases of averaging (sub)word embeddings from the last layer of neural Language Models, the results were either comparable or inferior to TF-IDF. The best-performing language models were GPT-1 and GPT-2. Fine-tuning of these on a subsample of a legal text corpus improved the results significantly, by a factor of 3–7 points. LMs seem to benefit neither from SIF nor from the removal of a single common component; their performance can, however, be mildly improved with a conventionally used decomposition, such as ICA (Hyvärinen and Oja, 2000).

Substantial improvement can be achieved by considering segments different from a single sentence, such as n-grams of sentences (meaning that any contiguous sequence of up to n sentences from a given text was scored and could be returned as a result).

Figure 2 presents how the performance of particular methods changes as a function of the number of example documents available within the simple similarity averaging scheme used in all the presented solutions. In general, the methods benefit substantially from the availability of a second example. A bigger number leads to a decreased variance but yields no improvement in the median score.

5 Discussion

The brief evaluation presented in the previous section has multiple limitations. First, it assumed retrieval of a single, most similar segment, whereas it appears that multiple clauses might be returned instead. However, we consider this restriction justifiable during a preliminary comparison of applicable methods. Multiple alternative selectors may be proposed in the future.

Secondly, all the evaluated methods assume scoring with the policy of averaging individual similarities. We encourage readers to experiment with different pooling methods or meta-learning strategies. Moreover, even the LM-based methods we had studied the most can be further studied in the proposed shared task. For example, only embeddings from the last layer were evaluated, even though it is possible that the higher layers may capture semantics better.

Finally, it is in principle possible to address the task in entirely different ways, for example, by per-

forming neither segmentation nor aggregation of word embeddings at all, but by matching clauses on the word level instead, which may be an interesting direction for further research. We decided to take the most common and straightforward way, due to fact performed evaluations are to serve as baselines for other methods.

6 Related Work

There is a large and varied body of work related to information retrieval in general; however, following Gillick et al. (2018) we consider the problem stated in an end-to-end manner, where the nearest neighbor search is performed on dense document representations. With this assumption, the main issue is to obtain reliable representations of documents, where by document we mean *any self-contained unit that can be returned to the user as a search result* (Büttcher et al., 2010). We use the term *segment* with the same meaning wherever it aids clarity.

Many approaches considered in the literature rely on word embedding and aggregation strategies. Simple methods proposed include averaging, as in the continuous bag-of-words (CBOW) model (Mikolov et al., 2013) or frequency-weighted averaging with the decomposition method applied (Arora et al., 2017). More sophisticated schemes include utilizing multiple weights, such as a novelty score, a significance score, and a corpus-wise uniqueness (Yang et al., 2018) or computing a vector of locally aggregated descriptors (Ionescu and Butnaru, 2019). Most of the proposed methods are orderless, and their limitations were recently discussed by Mai et al. (2019). However, there are also pooling approaches preserving spatial information, such as a hierarchical pooling operation (Shen et al., 2018). Other methods of obtaining sentence representations from word embeddings include training an autoencoder on a large collection of unlabeled data (Zhang et al., 2018) or utilizing random encoders (Wieting and Kiela, 2019). Despite its shortcomings and the availability of many sophisticated alternatives, the CBOW model is a common choice due to its ability to ensure strong results on many downstream tasks.

Different approaches assume training encoders through document embedding in an unsupervised or supervised manner, without the need for explicit aggregation. The former include Skip-Thought Vectors, trained with the objective of reconstructing the surrounding sentences of an encoded passage (Kiros et al., 2015). Although this method was outperformed by supervised models trained on a single NLI task (Conneau et al., 2017), paraphrase corpora (Jiao et al., 2018) or multiple tasks (Subramanian et al., 2018), the objective of predicting the next sentence is used as an additional objective in multiple novel models, such as the Universal Sentence Encoder (Cer et al., 2018). Even though many Transformer-based language models implement their own pooling strategy for generating sentence representations (special token pooling), they were shown to yield weak sentence embeddings, as described recently by Reimers and Gurevych (2019). The authors proposed a superior method of fine-tuning a pretrained BERT network with Siamese and triplet network structures to obtain sentence embeddings.

There were attempts to utilize semantic similarity methods explicitly in the legal domain, e.g., for a case law entailment within the COLIEE shared task. In a recent edition, Rabelo et al. (2019) used a BERT model fine-tuned on a provided training set in a supervised manner, and achieved the highest F-score among all teams. However, due to the reasons discussed in Section 4, their approach is not consistent with the nearest neighbor search, which is what we are aiming for.

7 Summary and Conclusions

We have introduced a new shared task of semantic retrieval from legal texts, which differs substantially from conventional NLI. It is heavily inspired by enterprise solutions referred to as *contract discovery*, focused on ensuring the inclusion of relevant clauses or their retrieval for further analysis. The main distinguishing characteristic of Contract Discovery shared task is conceptual, since:

- Candidate sequences are being mined from real texts. It is assumed span identification should be performed (systems should be able to return any document substring without any segmentation given in advance).
- It is suited for few-shot methods, filling the gap between conventional sentence classification and NLI tasks based on sentence pairs.

For the purposes of providing competetive baselines, we considered the problem stated in an end-to-end manner, where the nearest neighbor search is performed on document representations. With

this assumption, the main issue was to obtain representations of text fragments, which we referred to as segments. The description of the task was followed by the evaluation of multiple k-NN-based solutions within the unified framework, which may be used to describe future solutions. Moreover, a practical justification for handling the problem with k-NN was briefly introduced.

It has been shown that in this particular setting, pretrained, *universal* encoders fail to provide satisfactory results. One may suspect that this is a result of the difference between the domain they were trained on and the legal domain. During the evaluation, solutions based on the Language Models performed well, especially when unsupervised fine-tuning was applied. In addition to the aforementioned ability to fine-tune the method on legal texts, the most important indicator of success so far has been the involvement of multiple, sometimes overlapping substrings instead of sentences. Moreover, it has been demonstrated that the methods benefit substantially from the availability of a second example, and the presence of more leads to a decrease in variance, even when a simple similarity averaging scheme is applied.

The discussion regarding the presented methods and their limitations briefly outlined possible measures towards improving the baseline methods. In addition to the dataset and reference results, legal-specialized LMs have been made released to assist the research community in performing further experiments.

The Contract Discovery dataset, Edgar Corpus, we crawled, and all the mentioned models are publicly available on GitHub: `https://github.com/applicaai/contract-discovery`.

Acknowledgements

The Smart Growth Operational Programme supported this research under project no. POIR.01.01.01-00-0605/19 (*Disruptive adoption of Neural Language Modelling for automation of text-intensive work*).

There are no copyright issues regarding the Contract Discovery dataset, as both sources belong to the public domain. Documents were annotated ethically by our co-workers. Moreover, the colleagues who participated in annotation are among the authors of the paper.

References

Nada Almarwani, Hanan Aldarmaki, and Mona Diab. 2019. Efficient sentence embedding using discrete cosine transform.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot text classification with distributional signatures. *arXiv:1908.06039*.

Jason R. Baron, National Archives, Records Administration, and Office Of General. 2006. Trec-2006 legal track overview. In *In The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. 2012. *Assignment Problems. Revised reprint.* SIAM - Society of Industrial and Applied Mathematics. 393 Seiten.

Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. 2010. *Information Retrieval - Implementing and Evaluating Search Engines.* MIT Press.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Heting Chu. 2011. Factors affecting relevance judgment: a report from trec legal track. *Journal of Documentation*, 67:264–278.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, pages 993–1000, New York, NY, USA. ACM.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space.

Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.

N. Halko, P. G. Martinsson, and J. A. Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288.

Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13 4-5:411–30.

Radu Tudor Ionescu and Andrei M. Butnaru. 2019. Vector of Locally-Aggregated Word Embeddings (VLAWE): A Novel Document-level Representation.

Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`.

Xiaoqi Jiao, Fang Wang, and Dan Feng. 2018. Convolutional neural network for universal sentence embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2470–2481, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yoshinobu Kano, Mi Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of COLIEE 2017. In *COLIEE@ICAIL*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Florian Mai, Lukas Galke, and Ansgar Scherp. 2019. CBOW is not all you need: Combining CBOW with the compositional matrix space model. *CoRR*, abs/1902.06423.

K. Tamsin Maxwell and Burkhard Schafer. 2008. Concept and context in legal information retrieval. In *JURIX*.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Rashmi Nagpal, Chetna Wadhwa, Mallika Gupta, Samiulla Shaikh, Sameep Mehta, and Vikram Goyal. 2018. Extracting fairness policies from legal documents. *CoRR*, abs/1809.04262.

W. Douglas Oard, R. Jason Baron, Bruce Hedin, D. David Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for e-discovery. *Artif. Intell. Law*, pages 347–386.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *In EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005, Beijing, China. Coling 2010 Organizing Committee.

Juliano Rabelo, Mi-Young Kim, and Randy Goebel. 2019. Combining similarity and transformer methods for case law entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 290–296, New York, NY, USA. ACM.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *CoRR*, abs/1804.00079.

Javier Tejedor, Doroteo T. Toledano, Paula Lopez-Otero, Laura Docio-Fernandez, Mikel Peñagarikano, Luis Javier Rodriguez-Fuentes, and Antonio Moreno-Sandoval. 2019. Search on Speech from Spoken Queries: The Multi-Domain International ALBAYZIN 2018 Query-by-Example Spoken Term Detection Evaluation. *EURASIP J. Audio Speech Music Process.*, 2019(1).

Scott Vanderbeck, Joseph Bockhorst, and Chad Oldfather. 2011. A machine learning approach to identifying sections in legal briefs. In *MAICS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. 2019. Generalizing from a few examples: A survey on few-shot learning.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Ziyi Yang, Chenguang Zhu, and Weizhu Chen. 2018. Zero-training sentence embedding via orthogonal basis. *ArXiv*, abs/1810.00438.

Minghua Zhang, Yunfang Wu, Weikang Li, and Wei Li. 2018. Learning universal sentence representations with mean-max attention autoencoder.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

A File Structure

The documents' content can be found in the `reference.tsv` files. The input files `in.tsv` consist of tab-separated fields: Target ID (e.g. *57*), Clause considered (e.g. *governing-law*), Example #1 (e.g. *59 15215-15453*), . . . , Example #N. Each example consists of document ID and characters range. Ranges can be discontinuous. In such a case the sequences are separated with a comma, e.g. 4103-4882,12127-12971. The file with answers (`expected.tsv`) contains one answer per line, consisting of the entity name (to be copied from input) and characters range in the same format as described above. The reference file contains two tab-separated fields: document ID and content.

B Other Evaluation Results

Tables below describe evaluation results which were not included in the paper (or were included without broader context, that is without reference to different results from the same class of solutions).

Table 5 presents results with all the evaluated Sentence-BERT models. Table 6 shows scores achieved by TF-IDF with different settings, including other n-gram ranges. Results of particular Universal Sentence Encoder models are presented in Table 7. Table 8 shows results of Transformer-based Language Models not included in the paper. Finally, Table 9 is devoted to analysis of Discrete Cosine Transform embeddings.

Model	Soft F_1
bert-base-nli-cls-token	0.29
bert-base-nli-max-tokens	0.30
bert-base-nli-mean-tokens	0.31
bert-base-nli-stsb-mean-tokens	**0.32**
bert-base-wikipedia-sections-mean-tokens	0.25
bert-large-nli-cls-token	0.29
bert-large-nli-max-tokens	0.30
bert-large-nli-mean-tokens	0.30
bert-large-nli-stsb-mean-tokens	0.31
roberta-base-nli-mean-tokens	0.28
roberta-base-nli-stsb-mean-tokens	0.29
roberta-large-nli-mean-tokens	0.31
roberta-large-nli-stsb-mean-tokens	0.31

Table 5: Results of Sentence-BERT models on the *test-A* dataset when returning the most similar sentence. Names as in *sentence-transformers* library: https://github.com/UKPLab/sentence-transformers

Range (n-grams)	Binary	Soft F_1
1–1	–	0.32
1–2	–	0.35
1–3	–	0.36
1–1	+	0.36
1–2	+	**0.38**
1–3	+	0.37

Table 6: Results of TF-IDF on the *test-A* dataset when returning the most similar sentence.

Model	Soft F_1
multilingual/1	**0.38**
multilingual-large/1	0.33
multilingual-qa/1	0.28
large/3	0.26

Table 7: Results of Universal Sentence Encoder models on the *test-A* dataset when returning the most similar sentence.

Model	Soft F_1
bert-base-cased	0.25
bert-base-multilingual-cased	0.24
bert-base-multilingual-uncased	0.32
bert-base-uncased	0.26
bert-large-cased	0.21
bert-large-cased-whole-word-masking	0.31
bert-large-uncased	0.18
bert-large-uncased-whole-word-masking	0.35
roberta-base	0.25
roberta-large	0.32
openai-gpt	0.36
gpt2	0.16
gpt2-medium	0.11
gpt2-large	**0.41**

Table 8: Results of particular Transformer-based Language Models (without finetuning) on the *test-A* dataset when returning the most similar sentence. Names as in *transformers* library: https://github.com/huggingface/transformers

C	Soft F_1
c^0	**0.36**
$c^{0:1}$	0.30
$c^{0:2}$	0.25
$c^{0:3}$	0.20
$c^{0:4}$	0.18

Table 9: Results of GloVe embeddings (300d, EDGAR) on the *test-A* dataset when Discrete Cosine Transform sentence embeddings were created. The c^0 is equivalent to embeddings mean when k-NN methods are considered. The similar decrease of performance was observed for other models.

C Rest of the Clauses Considered

Random subsets of bond issue prospectuses and non-disclosure agreement documents from the US EDGAR database[6], as well as annual reports of charitable organizations from the UK Charity Register[7] were annotated, in such a way that clauses of the same type were selected (e.g. determining the governing law, merger restrictions, tax changes call or reserves policy). Clause types depend on the type of a legal act and can consist of a single sentence, multiple sentences or sentence fragments. Tables bellow present clause types annotated in each of the document groups.

Clause (Instances)	**Example**
GOVERNING LAW (152/160) The parties agree on which jurisdiction the contract will be subject to.	This Agreement shall be governed by and construed in accordance with the laws of the State of California without reference to its rules of conflicts of laws.
CONFIDENTIAL PERIOD (108/122) The parties undertake to maintain confidentiality for a certain period of time.	The term of this Agreement during which Confidential Information may be disclosed by one Party to the other Party shall begin on the Effective Date and end five (5) years after the Effective Date, unless extended by mutual agreement.
EFFECTIVE DATE (79/89) Information on the date of entry into force of the contract.	THIS AGREEMENT is entered into as of the 30th of July 2010 and shall be deemed to be effective as of July 23, 2010.
EFFECTIVE DATE REFERENCE (91/111)	This Contract shall become effective (the "Effective Date") upon the date this Contract is signed by both Parties.
NO SOLICITATION (101/117) Prohibition of acquiring employees of the other party (after the contract expires) and maintaining business relations with the customers of the other party.	You agree that for a period of eighteen months (18) from the date hereof you will not directly or indirectly recruit, solicit or hire any regional or district managers, corporate office employee, member of senior management of the Company (including store managers), or other employee of the Company identified to you.
CONFIDENTIAL INFORMATION FORM (152/174) Forms and methods of providing confidential information.	"Confidential Information" means any technical or commercial information or data, trade secrets, know-how, etc., of either Party or their respective Affiliates whether or not marked or stamped as confidential, including without limitation, Technology, Invention(s), Intellectual Property Rights, Independent Technology and any samples of products, materials or formulations including, without limitation, the chemical identity and any properties or specifications related to the foregoing. Any Development Program Technology, MPM Work Product, MSC Work Product, Hybrid Work Product, Prior End-Use Work Product and/or Shared Development Program Technology shall be Confidential Information of the Party that owns the subject matter under the terms set forth in this Agreement.
DISPUTE RESOLUTION (67/68) Arrangements for how to resolve disputes (arbitration, courts).	The Parties will attempt in good faith to resolve any dispute or claim arising out of or in relation to this Agreement through negotiations between a director of each of the Parties with authority to settle the relevant dispute. If the dispute cannot be settled amicably within fourteen (14) days from the date on which either Party has served written notice on the other of the dispute then the remaining provisions of this Clause shall apply.

Table 10: Clauses annotated in Non-disclosure Agreements. The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively.

[6] http://www.www.sec.gov/edgar.shtml
[7] http://www.gov.uk/find-charity-information

Clause (Instances)	Example
CHANGE OF CONTROL COVENANT (88/95) Information about the obligation to redeem bonds for 101% of the price in the event of change of control.	Upon the occurrence of a Change of Control Triggering Event (as defined below with respect to the notes of a series), unless we have exercised our right to redeem the notes of such series as described above under "Optional Redemption," the indenture provides that each holder of notes of such series will have the right to require us to repurchase all or a portion (equal to $2,000 or an integral multiple of $1,000 in excess thereof) of such holder's notes of such series pursuant to the offer described below (the "Change of Control Offer"), at a purchase price equal to 101% of the principal amount thereof, plus accrued and unpaid interest, if any, to the date of repurchase, subject to the rights of holders of notes of such series on the relevant record date to receive interest due on the relevant interest payment date.
CHANGE OF CONTROL NOTICE (78/79) Information about the obligation to inform bondholders (usually by mail) about the event of change of control. This clause usually follows immediately the above clause.	Within 30 days following any Change of Control, B&G Foods will mail a notice to each holder describing the transaction or transactions that constitute the Change of Control and offering to repurchase notes on the Change of Control Payment Date specified in the notice, which date will be no earlier than 30 days and no later than 60 days from the date such notice is mailed, pursuant to the procedures required by the indenture and described in such notice. Holders electing to have a note purchased pursuant to a Change of Control Offer will be required to surrender the note, with the form entitled "Option of Holder to Elect Purchase" on the reverse of the note completed, to the paying agent at the address specified in the notice of Change of Control Offer prior to the close of business on the third business day prior to the Change of Control Payment Date.
CROSS DEFAULT (96/110) The company does not comply with certain conditions (event of default), so the bonds become due (e.g. when the company does not submit financial statements on time) — our clause was limited to the event of non-repayment, usually the minimum sum is given.	due to our default, we (i) are bound to repay prematurely indebtedness for borrowed moneys with a total outstanding principal amount of $75,000,000 (or its equivalent in any other currency or currencies) or greater, (ii) have defaulted in the repayment of any such indebtedness at the later of its maturity or the expiration of any applicable grace period or (iii) have failed to pay when properly called on to do so any guarantee of any such indebtedness, and in any such case the acceleration, default or failure to pay is not being contested in good faith and not cured within 15 days of such acceleration, default or failure to pay;
LITIGATION DEFAULT (42/51) Court verdict or administrative decision which charge the company for a significant unpaid amount (another from the series of event of default).	(8) one or more judgments, orders or decrees of any court or regulatory or administrative agency of competent jurisdiction for the payment of money in excess of $30 million (or its foreign currency equivalent) in each case, either individually or in the aggregate, shall be entered against the Company or any subsidiary of the Company or any of their respective properties and shall not be discharged and there shall have been a period of 60 days after the date on which any period for appeal has expired and during which a stay of enforcement of such judgment, order or decree, shall not be in effect;
MERGER RESTRICTIONS (188/241) A clause preventing the merger or sale of a company, etc., except under certain conditions (generally, the company should not avoid its obligations to its bondholders).	Without the consent of the holders of the outstanding debt securities under the indentures, we may consolidate with or merge into, or convey, transfer or lease our properties and assets to any person and may permit any person to consolidate with or merge into us. However, in such event, any successor person must be a corporation, partnership, or trust organized and validly existing under the laws of any domestic jurisdiction and must assume our obligations on the debt securities and under the applicable indenture. We agree that after giving effect to the transaction, no event of default, and no event which, after notice or lapse of time or both, would become an event of default shall have occurred and be continuing and that certain other conditions are met; provided such provisions will not be applicable to the direct or indirect transfer of the stock, assets or liabilities of our subsidiaries to another of our direct or indirect subsidiaries. (Section 801)

BONDHOLDERS DEFAULT (191/241) A clause on the payment of the principal amount and interest — they become due as a result of an event of default, if such a declaration is made by bondholders.	If an event of default (other than an event of default referred to in clause (5) above with respect to us) occurs and is continuing, the trustee or the holders of at least 25% in aggregate principal amount of the outstanding notes by notice to us and the trustee may, and the trustee at the written request of such holders shall, declare the principal of and accrued and unpaid interest, if any, on all the notes to be due and payable. Upon such a declaration, such principal and accrued and unpaid interest will be due and payable immediately. If an event of default referred to in clause (5) above occurs with respect to us and is continuing, the principal of and accrued and unpaid interest on all the notes will become and be immediately due and payable without any declaration or other act on the part of the trustee or any holders.
TAX CHANGES CALL (48/56) A clause about the possibility of an earlier redemption of the bond by the issuer if the tax law or its interpretation changes.	If, as a result of any change in, or amendment to, the laws (or any regulations or rulings promulgated under the laws) of the Netherlands or the United States or any taxing authority thereof or therein, as applicable, or any change in, or amendments to, an official position regarding the application or interpretation of such laws, regulations or rulings, which change or amendment is announced or becomes effective on or after the date of the issuance of the notes, we become or, based upon a written opinion of independent counsel selected by us, will become obligated to pay additional amounts as described above in "Payment of additional amounts," then the Issuer may redeem the notes, in whole, but not in part, at 100% of the principal amount thereof together with unpaid interest as described in the accompanying prospectus under the caption "Description of WPC Finance Debt Securities and the Guarantee-Redemption for Tax Reasons."
FINANCIAL STATEMENTS (201/317) A clause on the obligation to submit (usually to the SEC) annual reports or other reports.	Notwithstanding that the Company may not be subject to the reporting requirements of Section 13 or 15(d) of the Exchange Act, the Company will file with the SEC and provide the Trustee and Holders and prospective Holders (upon request) within 15 days after it files them with the SEC, copies of its annual report and the information, documents and other reports that are specified in Sections 13 and 15(d) of the Exchange Act. In addition, the Company shall furnish to the Trustee and the Holders, promptly upon their becoming available, copies of the annual report to shareholders and any other information provided by the Company to its public shareholders generally. The Company also will comply with the other provisions of Section 314(a) of the TIA.

Table 11: Clauses annotated in Corporate Bonds. The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively.