



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

Łukasz Borchmann

# 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 Szalkiewicz  
Gabriela Pałka and Karol Kaczmarek and Agnieszka Kaliska and Filip Graliński

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

## Abstract

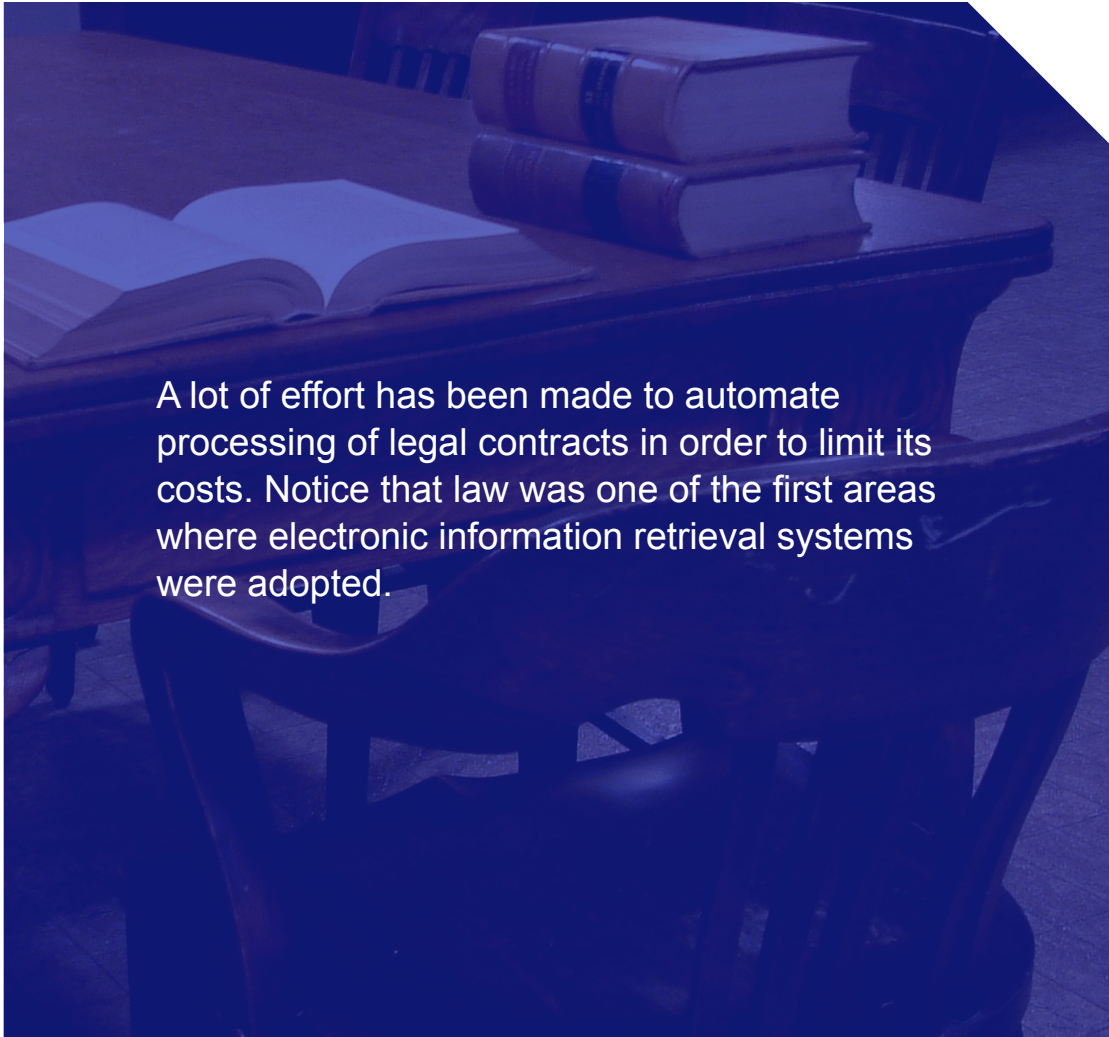
We propose a new shared task of semantic retrieval from legal texts, in which a so-called *contract discovery* is to be performed—where legal clauses are extracted from documents, given a few examples of similar clauses from other legal acts. The task differs substantially from conventional NLI and shared 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 frame-

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*). Most of the few-shot tasks do not include the

# Agenda

- 01. Problem statement, examples
- 02. Desiderata
- 03. Comparison to existing resources
- 04. Creation of Contract Discovery dataset
- 05. Baselines
- 06. Future directions
- 07. Few of our other works



A lot of effort has been made to automate processing of legal contracts in order to limit its costs. Notice that law was one of the first areas where electronic information retrieval systems were adopted.

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.

One may expect only a minimal number of examples in a typical business case, because the process is performed constantly for different clauses, and it is practically impossible to prepare data in a number required by a conventional classifier every time.



# Hot example



## Situation

Support for the London Interbank Offered Rate (LIBOR) benchmark is set to expire by 2021.

LIBOR has been the most widely used interest rate in the world; it has been used to price assets worth over \$370 trillion.



## Problem

Many institutions face the risk that millions of contracts – even hundreds of millions of contracts – will wind up invalid or exposed to risk because the interest rates they reference are formulated in terms of a benchmark no longer in play.



## Action required

There is a need to locate as many as a few dozen passages from each contract in question, with particular attention to any provisions that constitute a so-called fallback clause relevant to the replacement of LIBOR.

# Other

## 01. Reserves policy

What are the current financial reserves of the organization and how much these reserves should be as assumed?

## 03. Auditor opinion

Summary of the opinion of an independent auditor or inspector, often in the form of a list of points.

## 05. Effective date

Information on the date of entry into force of the contract.

## 07. Merger restrictions

A clause preventing the merger or sale of a company, etc., except under certain conditions.

## 02. Governing law

The parties agree on which jurisdiction the contract will be subject to.

## 04. Confidential period

The parties undertake to maintain confidentiality for a certain period of time.

## 06. No solicitation

Prohibition of acquiring employees of the other party and maintaining business relations with the customers of the other party.

## 08. Litigation default

Court verdict or administrative decision which charge the company for a significant unpaid amount (another from the series of event of default).

# Desiderata

We wished to construct a dataset for testing the mechanisms that detect various types of regulations in legal documents.



## No formal structure

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.



## Span identification

It is assumed that a searched passage can be any part of the document and not necessarily a complete paragraph, subparagraph, or a clause.



## No question required

We intended to use a query-by-example scenario instead of the setting where articles are being returned as an answer for the question specified in natural language.



## Few-shot evaluation

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.

# Towards dataset and shared task

600 acts from 3 classes

21 clause types

2,500 spans

Few-shot

## Documents

Random subsets of bond issue prospectuses and non-disclosure agreement documents from the US EDGAR database, as well as annual reports of charitable organizations from the UK Charity Register were sampled.

## Types determination

Clause types depend on the type of a legal act and can consist of a single sentence, multiple sentences or sentence fragments.

We restricted ourselves to 21 types as a result of a trade-off between annotation cost and the ability to formulate general remarks.

## Annotation

Each document was annotated by two experts, and then reviewed by a super-annotator, who decided the gold standard.

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).

## Evaluation procedure

Evaluation is performed by means of a repeated random sub-sampling validation procedure.

We few-shot learning with 1-5 documents available. Soft F1 metric on character-level spans is used for the purpose of evaluation.



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

## Comparison

None of existing tasks involving semantic similarity or legal information extraction, assume span identification.

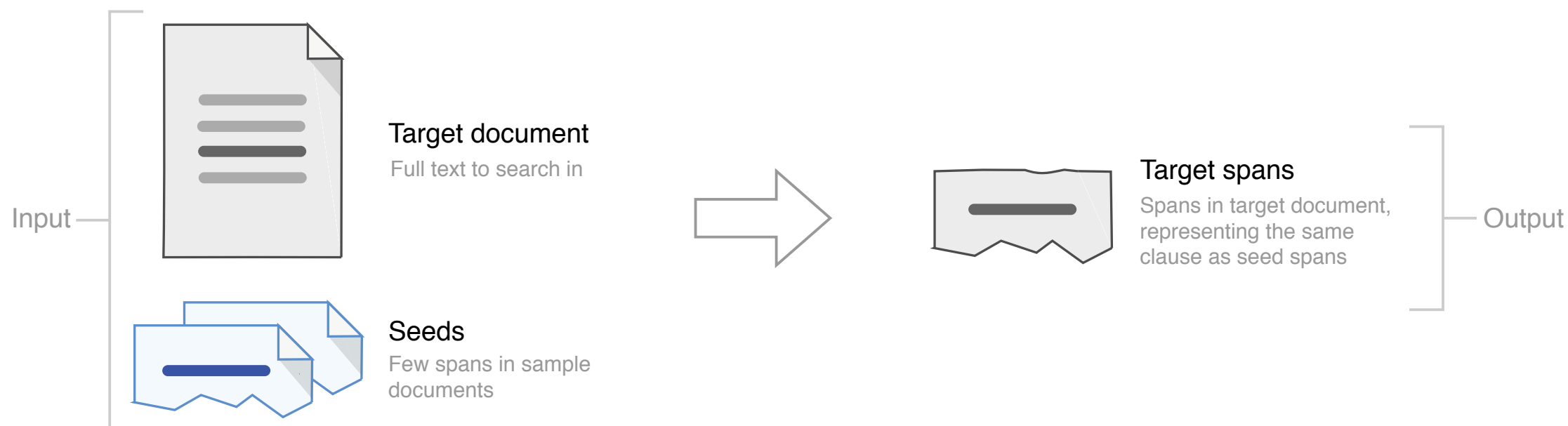
There are a few NLP tasks where span identification is performed, e.g., some of plagiarism detection competitions and recently introduced SemEval task.

When different media are considered, it is equivalent to the action recognition in temporally untrimmed videos or query-by-example spoken term detection.

Moreover, the few-shot setting is not popular within the field of NLP yet.

## Task recap

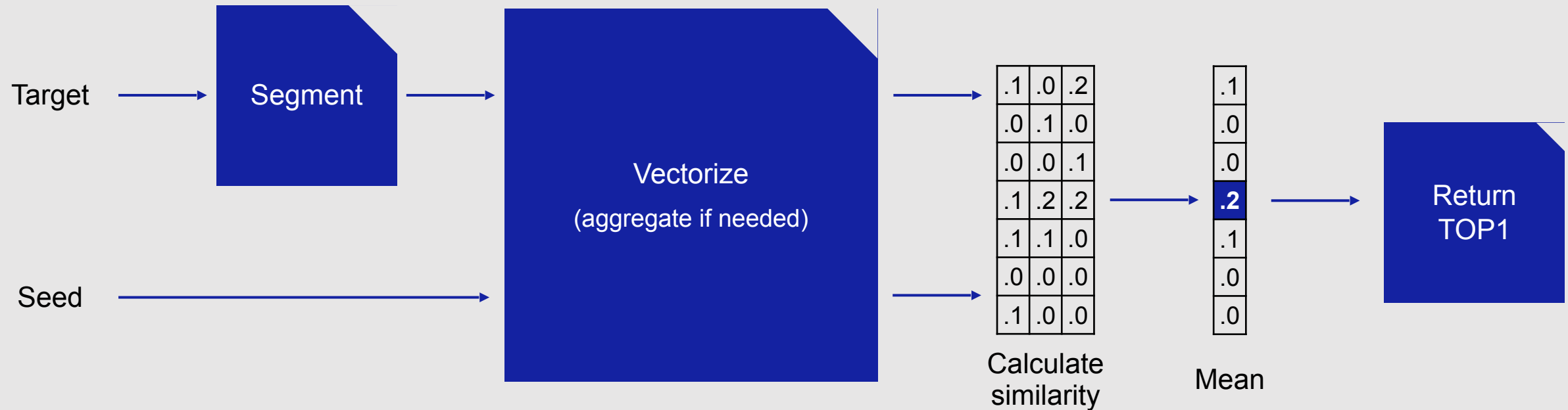
The aim of our task is to identify spans in the requested documents representing clauses analogous to the spans selected in other documents.



# How to solve this problem?

## Baselines

We proposed a simple k-NN-based approach as a reference. It assumes segmentation of target documents, pre-encoding all candidate segments, and returning a single segment with the highest mean cosine similarity to seed examples.





## TF-IDF and static word embeddings

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.

Segmenter	Vectorizer	Projector	Scorer	Aggregator	Soft F1
sentence	TF-IDF (1–2 grams, binary TF term)	—	mean cosine	—	0.38
		tSVD	mean cosine	—	0.39
sentence	GloVe (300d, Wikipedia & Gigaword)	—	mean cosine	mean	0.34
		—	mean WMD	—	0.35
		SIF SVD	mean cosine	SIF	0.37
sentence	GloVe (300d, EDGAR)	—	mean cosine	mean	0.36
		—	mean WMD	—	0.35
		SIF SVD	mean cosine	SIF	<b>0.41</b>

TF-IDF with truncated SVD decomposition is commonly referred to as Latent Semantic Analysis.

SVD in SIF method is used to perform removal of single common component.

## Language models and sentence encoders

Sentence-BERT and Universal Sentence Encoder could not outperform the simple TF-IDF approach. 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.

Segmenter	Vectorizer	Projector	Scorer	Aggregator	Soft F1
sentence	Sentence-BERT	—	mean cosine	mean	0.32
	USE	—	mean cosine	—	0.38
sentence	BERT	—	mean cosine	mean	0.35
	GPT-1	—	mean cosine	—	0.36
	GPT-2	—	mean cosine	—	<b>0.41</b>
	RoBERTa	—	mean cosine	—	0.31

Only the best models from each architecture are presented here.

## Fine-tuning and n-grams of sentences

Fine-tuning on a subsample of a legal text corpus improved the results significantly. 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.

Substantial improvement can be achieved by considering segments different from a single sentence, such as n-grams of sentences.

Segmenter	Vectorizer	Projector	Scorer	Aggregator	Soft F1
sentence	fine-tuned GPT-1	—	mean cosine	mean	0.43
	fine-tuned GPT-1	fICA	mean cosine	mean	0.44
	fine-tuned GPT-2	—	mean cosine	mean	0.44
	fine-tuned GPT-2	fICA	mean cosine	mean	0.45
1-3 sen.	fine-tuned GPT-1	—	mean cosine	mean	0.47
	fine-tuned GPT-1	fICA	mean cosine	mean	0.49
	fine-tuned GPT-2	—	mean cosine	mean	0.46
	fine-tuned GPT-2	fICA	mean cosine	mean	<b>0.51</b>
human					<b>0.84</b>

Range in segmented means that any contiguous sequence of up to n sentences from a given text was scored and could be returned as a result.

ICA can be viewed as a generalization of PCA to non-gaussian data (See Ivanov, 2017).

# Future

## 01. Multiple results

Baselines assume retrieval of a single, most similar segment, whereas it appears that multiple clauses might be returned instead.

## 03. Meta-learning

The problem cannot be solved by means of conventional classifiers although one can attempt it with, e.g., the prototypical network.

## 05. Idea of yours

<https://github.com/applicaai/contract-discovery>

## 02. Scoring policy

All the evaluated methods assume scoring with the policy of averaging individual similarities.

## 04. Word-level

Performing neither segmentation nor aggregation of word embeddings at all, but by matching clauses on the word level instead.




# Few of our other recent works

Dynamic Boundary Time Warping for Sub-sequence matching with Few Examples (ESWA)

From Dataset Recycling to Multi-Property Extraction and Beyond (CoNLL)

ARTICLE IN PRESS

Expert Systems With Applications xxx (xxxx) xxx




ELSEVIER

Contents lists available at [ScienceDirect](#)

Expert Systems With Applications

journal homepage: [www.elsevier.com/locate/eswa](http://www.elsevier.com/locate/eswa)



### Dynamic Boundary Time Warping for sub-sequence matching with few examples

Lukasz Borchmann<sup>a,\*</sup>, Dawid Jurkiewicz<sup>a,1</sup>, Filip Graliński<sup>a</sup>, Tomasz Górecki<sup>b</sup>

<sup>a</sup> *Applica.ai, Warsaw, Poland*  
<sup>b</sup> *Adam Mickiewicz University, Poznań, Poland*

ARTICLE INFO

ABSTRACT

**Keywords:**  
Dynamic Time Warping  
Sub-sequence matching  
Natural Language Processing  
Information Retrieval  
Few-shot learning  
Semantic retrieval

The paper presents a novel method of finding a fragment in a long temporal sequence similar to the set of shorter sequences. We are the first to propose an algorithm for such a search that does not rely on computing the average sequence from query examples. Instead, we use query examples as is, utilizing all of them simultaneously. The introduced method based on the Dynamic Time Warping (DTW) technique is suited explicitly for few-shot query-by-example retrieval tasks. We evaluate it on two different few-shot problems from the field of Natural Language Processing. The results show it either outperforms baselines and previous approaches or achieves comparable results when a low number of examples is available.

### From Dataset Recycling to Multi-Property Extraction and Beyond

Tomasz Dwojak<sup>1,2</sup>, Michał Pietruszka<sup>1,3</sup>, Łukasz Borchmann<sup>1,4</sup>,  
Jakub Chłędowski<sup>1,3</sup>, and Filip Graliński<sup>1,2</sup>

<sup>1</sup>Applica.ai  
<sup>2</sup>Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań  
<sup>3</sup>Faculty of Mathematics and Computer Science, Jagiellonian University  
<sup>4</sup>Institute of Computing Science, Poznań University of Technology

[tomasz.dwojak@applica.ai](mailto:tomasz.dwojak@applica.ai)

Abstract

This paper investigates various Transformer architectures on the WikiReading Information Extraction and Machine Reading Comprehension dataset. The proposed dual-source model outperforms the current state-of-the-art by a large margin. Next, we introduce WikiReading Recycled—a newly developed public dataset, and the task of multiple-property extraction. It uses the same data as WikiReading but does not inherit its predecessor's identified disadvantages. In addition, we provide a human-annotated test set with diagnostic subsets for a detailed analysis of model

QNLI (Raffel et al., 2020), CoLA or MRPC (Wang et al., 2020). However, as a consequence of different kinds of noise in the data, they rarely maximize the score metric (Stanisławek et al., 2019). While current work in NLP is focused on preparing new datasets, we regard recycling the current ones as equally important as creating a new one. Thus, after outperforming previous state-of-the-art on WikiReading, we investigated the dataset's weaknesses and created an entirely new, more challenging Multi-Property Extraction task with improved data splits and a reliable, human-annotated test set.

## Few of our other recent works

LAMBERT: Layout-Aware language Modeling using BERT for information extraction

On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them (SemEval @ COLING)

### LAMBERT: Layout-Aware (Language) Modeling using BERT for information extraction

Łukasz Garncarek\*   Rafał Powalski\*   Tomasz Stanisławek\*  
Bartosz Topolski\*   Piotr Halama  
Applica.ai, ul. Zajęcza 15, 00-351 Warszawa (Poland)  
firstname.lastname@applica.ai

Filip Graliński  
Applica.ai, ul. Zajęcza 15, 00-351 Warszawa (Poland)  
Adam Mickiewicz University,  
ul. Wieniawskiego 1, 61-712 Poznań (Poland)  
filip.gralinski@applica.ai

April 29, 2020

#### Abstract

In this paper we introduce a novel approach to the problem of understanding documents where the local semantics is influenced by non-trivial layout. Namely, we modify the Transformer architecture in a way that allows it to use the graphical features defined by the layout, without the need to re-learn the language semantics from scratch, thanks to starting the training process from a model pretrained on classical language modeling tasks.

### ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them

Dawid Jurkiewicz\*   Łukasz Borchmann\*   Izabela Kosmala   Filip Graliński

Applica.ai   Zajęcza 15, 00-351 Warsaw, Poland  
firstname.lastname@applica.ai

#### Abstract

This paper presents the winning system for the propaganda Technique Classification (TC) task and the second-placed system for the propaganda Span Identification (SI) task. The purpose of the TC task was to identify an applied propaganda technique given propaganda text fragment. The goal of SI task was to find specific text fragments which contain at least one propaganda technique. Both of the developed solutions used semi-supervised learning technique of self-training. Interestingly, although CRF is barely used with transformer-based language models, the SI task was approached with RoBERTa-CRF architecture. An ensemble of RoBERTa-based models was proposed for the TC task, with one of them making use of Span CLS layers we introduce in the present paper. In addition to describing the submitted systems, an impact of architectural decisions and training schemes is investigated along with remarks regarding training models of the same or better quality with lower computational budget. Finally, the results of error analysis are presented.

#### 1 Introduction

# Thank you