Neural Contract Element Extraction Revisited

Ilias Chalkidis, **Manos Fergadiotis**, **Prodromos Malakasiotis**, and
Ion Androutsopoulos

Institute of Informatics & Telecommunications, NCSR "Demokritos", Greece
Department of Informatics, Athens University of Economics and Business, Greece

Abstract

We investigate contract element extraction. We show that LSTM-based encoders perform better than dilated CNNs, Transformers, and BERT in this task. We also find that domain-specific WORD2VEC embeddings outperform generic pre-trained GLOVE embeddings. Morpho-syntactic features in the form of POS tag and token shape embeddings, as well as context-aware ELMO embeddings do not improve performance. Several of these observations contradict choices or findings of previous work on contract element extraction and generic sequence labeling tasks, indicating that contract element extraction requires careful task-specific choices.

1 Introduction

Extracting information from contracts and other legal agreements is an important part of daily business worldwide. Thousands of agreements are written up every day, resulting in a huge volume of legal documents relating to several business processes, such as employment, services/vendors, loans, leases, investments. These documents contain crucial information (e.g., contract terms, pay rates, termination rights). More importantly, when negotiating or revising agreements, the parties involved need to scrutinize all the terms of the agreements as recorded in the corresponding documents.

In this work, we focus on contract element extraction, i.e., extracting core information (e.g., parties, dates of interest, means of dispute, amounts) from contracts. Following Chalkidis et al. [1, 2], the task is viewed as sequence labeling, i.e., we aim to classify each token as (possibly part of) a party name, effective date, termination date, jurisdiction, address, amount etc. or 'none'. However, we use a single (multi-class) classifier for all the contract elements that may reside in each contract zone (e.g., header or applicable law section), which allows the classifier of each zone to generalize across contract entity types, whereas Chalkidis et al. used a separate (binary) classifier for each contract element type per zone. Furthermore, we investigate how the following three factors affect the extractors (classifiers).

Sequence encoders: We compare several neural encoders, namely stacked BILSTMs [5], DILATED-CNNs [7], stacked TRANSFORMERs [14], and BERT [4], whereas Chalkidis et al. [1] considered only stacked BILSTMs. Contrary to previous studies on generic sequence labeling tasks [13], we show that DILATED-CNNs are not comparable to stacked BILSTMs. More interestingly, we show that stacked BILSTMs also outperform the state-of-the-art language representation model BERT in this task.

CRF layers: We show that the use of (linear-chain) CRFs [9] on top of each encoder has a significant positive impact in all encoders, contrary to the findings of Chalkidis et al. [1], where the contribution of the CRF layer was unclear. This is most probably due to our use of multi-class classifiers, which leads to more constraints in the permissible sequences of predicted labels.

Input representations: We experiment with 200-D GLOVE word embeddings and domain-specific 200-D WORD2VEC embeddings pre-trained on approx. 750k contracts [2].[1] We also consider

[1] We used 200-D WORD2VEC embeddings, as Chalkidis et al., and generic GLOVE embeddings that are available in the same dimensionality (trained on 6 billions tokens from Wikipedia and Gigaword).

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

additional WORD2VEC embeddings representing POS tags and token shapes [1], character-based word embeddings obtained by character-level CNNs [11], and context-aware ELMO embeddings [12]. Domain-specific WORD2VEC embeddings outperform generic GLOVE embeddings. Morpho-syntactic (POS tags, token shapes), character-level, and context-aware embeddings (ELMO) increase computational complexity without delivering any significant performance improvement.

2 Related Work

Huang et al. [6] introduced BILSTM-CRF for sequence labeling (e.g., NER, POS tagging). Lample et al. [10] and Ma et al. [11] further improved the performance of BILSTM-CRF by adding character-based word embeddings obtained with BILSTM and CNN character encoders. Chiu et al. [3] reported mixed results by adding capitalization features and gazetteers. Strubell et al. [13] reported that DILATED-CNNs [7] have comparable results with BILSTMs, while being faster. Peters et al. [12] reported gains on several datasets, including the CONLL-2003 NER dataset, by exploiting context-aware word representations (ELMO), while Devlin et al. [4] achieved further improvements with BERT. Chalkidis et al. [2] introduced the task of contract element extraction, initially showing that linear window-based classifiers outperform rule-based ones. Later they improved performance in most cases using LSTM-based methods [1], with BILSTM-CRF being one of the best. They did not, however, compare to alternative (other than LSTM-based) encoders, neither did they investigate the necessity of the morpho-syntactic features they included in their input representations. They also did not experiment with character-based word representations and context-aware word embeddings (ELMO).

3 Task Definition and Datasets

We experimented with two subsets (Header/Preamble and Applicable Law) of the publicly available data provided by Chalkidis et al. [1, 2], and an in-house dataset with sections from lease agreements.

Contract Header / Preamble This subset contains the contract headers of the contracts of Chalkidis et al., where the goal is to identify *contract titles* (3836 training/650 test element instances), *parties* (6780/1250), *start dates* (2210/293) and *effective dates* (594/85).

Applicable Law : This subset contains the sections of the contracts of Chalkidis et al. where the *governing law* (2080 training/289 test) and *jurisdiction* (1245/229) elements need to be identified.

Lease Particulars : This dataset contains sections from lease agreements with the following elements: address of the leased *property* (2066 training/486 test), *landlord* (2269/559), *tenant* (2110/519), *start date* (1458/346), *effective date* (971/248), *end date* (821/216), *term* (period) of the lease (869/196), *rent amount* (1776/457).

4 Experiments

Experimental Setup: We used HYPEROPT and 5-fold Monte Carlo cross-validation to tune the following hyper-parameters on the training data with the following ranges: ENCODER output units {100, 150, 200, 250, 300}, ENCODER layers {1, 2, 3, 4}, batch size {8, 12, 16, 24, 32}, DROPOUT rate {0.2, 0.3, 0.4, 0.5, 0.6}, word DROPOUT rate {0.0, 0.05, 0.1}. All models were evaluated in precision, recall and F1-score per entity (entire element, *strict*) as defined in SemEval-2013 Task 9.1.[2] We use the ADAM optimizer [8] with initial learning rate 0.001.

Alternative Encoders: Table 1 reports results with different sequence encoders, always followed by a CRF layer. In these experiments except for BERT-CRF, the input representation of each token is the concatenation of its word, POS, and shape WORD2VEC embeddings, as in Chalkidis et al. [1]. With all encoders, a dense layer with a softmax activation operates on the top-level representation of each token, providing a probability distribution over the labels, which is fed to the CRF. Contrary to recent findings in sequence labeling [13], BILSTMs outperform DILATED-CNNs, stacked TRANSFORMERs, and BERT in all cases.[3] We can only speculate that although TRANSFORMERs and BERT include

[2]For details on the evaluation rules, see `https://www.cs.york.ac.uk/semeval-2013/task9/data/uploads/semeval_2013-task-9_1-evaluation-metrics.pdf`. We report mean scores on test data.

[3]We use the publicly available BERT-BASE model of Devlin et al. [4], fine-tuned on our training data.

	BILSTM-CRF			DILATED-CNNS-CRF			TRANSFORMERS-CRF			BERT-CRF		
CONTRACT HEADER												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Title	**96.0**	**96.4**	**96.2**	94.7	94.9	94.8	93.1	93.2	93.1	92.4	93.4	92.9
Party	**95.3**	**88.9**	**92.0**	93.7	86.2	89.8	88.4	79.4	83.6	87.3	85.4	86.3
S. Date	**96.8**	**97.4**	**97.1**	91.3	96.6	93.8	91.3	92.7	92.0	94.3	95.7	95.0
E. Date	94.6	**96.9**	95.7	**96.9**	95.1	**95.9**	92.0	88.5	90.1	83.9	84.7	84.3
MACRO-AVG	**95.7**	**94.9**	**95.2**	94.1	93.2	93.6	91.2	88.4	89.7	89.5	89.8	89.6
APPLICABLE LAW												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Jurisdiction	**79.7**	**72.4**	**75.9**	69.6	67.6	68.4	73.6	58.5	65.0	73.8	67.3	70.4
Gov. Law	**98.1**	**96.3**	**97.2**	95.1	92.5	93.8	98.0	90.3	94.0	92.7	91.9	92.3
MACRO-AVG	**88.9**	**84.4**	**86.5**	82.3	80.0	81.1	85.8	74.4	79.5	83.2	79.6	81.3
LEASE PARTICULARS												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Property	**67.0**	**65.8**	**66.2**	61.8	61.8	61.7	53.9	50.1	51.8	51.6	52.8	52.1
Landlord	**87.7**	**86.6**	**87.2**	83.4	83.8	83.6	76.5	68.7	72.3	76.4	81.8	78.8
Tenant	**90.7**	**90.9**	**90.8**	89.7	87.8	88.7	81.5	72.4	76.6	84.8	87.3	86.0
S. Date	**92.4**	**95.0**	**93.7**	91.7	93.4	92.5	88.2	90.5	89.3	89.0	94.1	91.4
E. Date	**88.7**	**90.8**	**89.7**	81.1	87.5	84.1	79.9	71.1	75.2	87.6	89.2	88.3
T. Date	**93.9**	**85.4**	**89.3**	91.3	84.2	87.6	73.2	67.7	70.2	85.6	88.7	87.1
Term	**86.6**	**89.1**	**87.8**	81.9	87.0	84.3	76.5	75.5	75.8	80.3	86.2	83.1
Rent	**86.5**	**86.0**	**86.2**	81.2	82.5	81.7	81.4	74.3	77.5	76.8	91.1	83.3
MACRO-AVG	**86.7**	**86.2**	**86.4**	82.8	83.5	83.0	76.4	71.3	73.6	79.0	83.9	81.3

Table 1: Results with alternative sequence encoders. BILSTM-based models are clearly better.

positional embeddings and have large receptive fields (via stacking), BILSTM-based models still cope better with long-term dependencies, which are important in legal documents. For example, to distinguish start and effective dates, or tenants, landlords and other parties (e.g., guarantors, representatives), one often has to consider a broader context than in generic named entity recognition.

Impact of CRFs: Table 2 compares the performance of all encoders with and without CRFs. In each dataset, results are macro-averaged over contract element types. Similarly to prior sequence labeling studies [10, 13] and unlike Chalkidis et al. [1], we find that CRFs almost always improve performance, especially for non-BILSTM encoders. The only exception is BERT in CONTRACT HEADER.

	CONTRACT HEADER			APPLICABLE LAW			LEASE PARTICULARS		
	P	R	F1	P	R	F1	P	R	F1
BILSTMs	93.4	94.0	93.7	81.6	80.7	81.1	82.0	82.7	82.3
+ CRF	**95.7**	**94.9**	**95.2**	**88.9**	**84.4**	**86.5**	**86.7**	**86.2**	**86.4**
DILATED-CNNs	84.2	88.0	86.0	68.7	72.7	70.5	65.9	74.3	69.8
+ CRF	94.1	93.2	93.6	82.3	80.0	81.1	82.8	83.5	83.0
TRANSFORMERs	81.8	86.4	84.0	54.5	53.9	54.1	58.0	64.1	60.8
+ CRF	91.2	88.4	89.7	85.8	74.4	79.5	76.4	71.3	73.6
BERT	90.0	90.9	90.4	78.3	78.1	78.2	77.0	79.8	78.2
+ CRF	89.5	89.8	89.6	83.2	79.6	81.3	79.0	83.9	81.3

Table 2: Macro-averaged results with/without CRF layers. CRFs almost always improve performance.

Alternative Feature Representations: Table 3 compares the performance of BILSTM-CRF, the best encoder, with different input representations. Generic word embeddings (GLOVE) are vastly outperformed by domain-specific ones (W2V-WORD). Adding POS tag and token shape embeddings (W2V-ALL) does not improve overall performance (see F1 scores). Adding character-level word embeddings (W2V-WORD+CHAR) also has no consistent or significant positive impact on F1. ELMO embeddings also do not lead to consistent noticeable improvement, possibly because the generic corpora (and word contexts) that ELMO was trained on are very different than contracts.

	CONTRACT HEADER			APPLICABLE LAW			LEASE PARTICULARS		
	P	R	F1	P	R	F1	P	R	F1
GLOVE (generic)	90.2	89.6	89.9	88.7	84.3	86.4	66.1	65.8	65.9
W2V-WORD (domain-specific)	95.7	**95.1**	**95.4**	89.0	84.1	86.5	87.0	86.2	86.6
W2V-ALL (incl. POS, shape)	95.7	94.9	95.2	88.9	**84.4**	86.5	86.7	86.2	86.4
W2V-ALL+CHAR	**96.1**	94.0	95.0	**89.3**	82.2	85.5	**87.8**	86.1	**86.9**
W2V-ALL+ELMO	95.8	94.8	95.3	**89.3**	84.2	**86.7**	86.0	**87.5**	86.7

Table 3: Macro-averaged BILSTM-CRF results with alternative input representations. Domain-specific WORD2VEC embeddings outperform generic GLOVE ones. POS and token shape (ALL), character-based (CHAR), and context-aware embeddings (ELMO) lead to no significant/consistent improvement.

5 Limitations and Future Work

ELMO and BERT were pre-trained on generic corpora, like the English Wikipedia and BookCorpus. Pre-training (or further training) them on contractual corpora will improve their performance, possibly surpassing our currently best methods. We leave this for future work and further investigation of the superior results of BILSTM-based methods on our datasets.

References

[1] I. Chalkidis and I. Androutsopoulos. A Deep Learning Approach to Contract Element Extraction. In *JURIX*, Luxembourg, 2017.

[2] I. Chalkidis, I. Androutsopoulos, and A. Michos. Extracting Contract Elements. In *ICAIL*, London, UK, 2017.

[3] J. Chiu and E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[4] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, Minneapolis, USA, 2018.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[6] Z. Huang, W. Xu, and K. Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *Arxiv*, 2015.

[7] N. Kalchbrenner, L. Espeholt, K. Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural Machine Translation in Linear Time. *Arxiv*, 2016.

[8] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, San Diego, USA, 2015.

[9] J.D. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, Williamstown, MA, USA, 2001.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural Architectures for Named Entity Recognition. In *ACL*, San Diego, USA, 2016.

[11] X. Ma and E. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *ACL*, Berlin, Germany, 2016.

[12] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. In *NAACL-HLT*, New Orleans, USA, 2018.

[13] E. Strubell, P. Verga, D. Belanger, and A. McCallum. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *EMNLP*, Copenhagen, Denmark, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In *NeurIPS*, Long Beach, USA, 2017.