Simple BERT Models for Relation Extraction and Semantic Role Labeling

Peng Shi and **Jimmy Lin**
David R. Cheriton School of Computer Science
University of Waterloo
{peng.shi, jimmylin}@uwaterloo.ca

Abstract

We present simple BERT-based models for relation extraction and semantic role labeling. In recent years, state-of-the-art performance has been achieved using neural models by incorporating lexical and syntactic features such as part-of-speech tags and dependency trees. In this paper, extensive experiments on datasets for these two tasks show that without using any external features, a simple BERT-based model can achieve state-of-the-art performance. To our knowledge, we are the first to successfully apply BERT in this manner. Our models provide strong baselines for future research.

1 Introduction

Relation extraction and semantic role labeling (SRL) are two fundamental tasks in natural language understanding. The task of relation extraction is to discern whether a relation exists between two entities in a sentence. For example, in the sentence "Obama was born in Honolulu", "Obama" is the subject entity and "Honolulu" is the object entity. The task of a relation extraction model is to identify the relation between the entities, which is `per:city_of_birth` (birth city for a person). For SRL, the task is to extract the predicate–argument structure of a sentence, determining "who did what to whom", "when", "where", etc. Both capabilities are useful in several downstream tasks such as question answering (Shen and Lapata, 2007) and open information extraction (Fader et al., 2011).

State-of-the-art neural models for both tasks typically rely on lexical and syntactic features, such as part-of-speech tags (Marcheggiani et al., 2017), syntactic trees (Roth and Lapata, 2016; Zhang et al., 2018; Li et al., 2018), and global decoding constraints (Li et al., 2019). In particular, Roth and Lapata (2016) argue that syntactic features are necessary to achieve competitive performance in dependency-based SRL. Zhang et al. (2018) also showed that dependency tree features can further improve relation extraction performance. Although syntactic features are no doubt helpful, a known challenge is that parsers are not available for every language, and even when available, they may not be sufficiently robust, especially for out-of-domain text, which may even hurt performance (He et al., 2017).

Recently, the NLP community has seen excitement around neural models that make heavy use of pretraining based on language modeling (Peters et al., 2018; Radford et al., 2018). The latest development is BERT (Devlin et al., 2018), which has shown impressive gains in a wide variety of natural language tasks ranging from sentence classification to sequence labeling. A natural question follows: can we leverage these pretrained models to further push the state of the art in relation extraction and semantic role labeling, without relying on lexical or syntactic features? The answer is yes. We show that simple neural architectures built on top of BERT yields state-of-the-art performance on a variety of benchmark datasets for these two tasks. The remainder of this paper describes our models and experimental results for relation extraction and semantic role labeling in turn.

2 BERT for Relation Extraction

2.1 Model

For relation extraction, the task is to predict the relation between two entities, given a sentence and two non-overlapping entity spans. In order to encode the sentence in an entity-aware manner, we propose the BERT-based model shown in Figure 1. First, we construct the input sequence [[CLS] sentence [SEP] subject [SEP] object [SEP]]. To prevent overfitting, we replace the entity mentions in

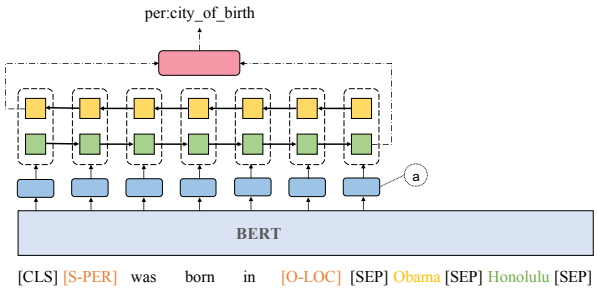

Figure 1: Architecture of our relation extraction model. **(a)** denotes the concatenation of BERT contextual embedding and position embedding. The final prediction is based on the concatenation of the final hidden state in each direction from the BiLSTM, fed through an MLP.

the sentence with masks, comprised of argument type (subject or object) and entity type (such as location and person), e.g., SUBJ-LOC, denoting that the subject entity is a location.

The input is then tokenized by the WordPiece tokenizer (Sennrich et al., 2016) and fed into the BERT encoder. After obtaining the contextual representation, we discard the sequence after the first [SEP] for the following operations.

We use $\mathcal{H} = [h_0, h_1, ..., h_n, h_{n+1}]$ to denote the BERT contextual representation for [[CLS] sentence [SEP]]. Note that n can be different from the length of the sentence because the tokenizer might split words into sub-tokens. The subject entity span is denoted $\mathcal{H}_s = [h_{s_1}, h_{s_1+1}, ..., h_{s_2}]$ and similarly the object entity span is $\mathcal{H}_o = [h_{o_1}, h_{o_1+1}, ..., h_{o_2}]$. Following Zhang et al. (2017), we define a position sequence relative to the subject entity span $[p_0^s, ..., p_{n+1}^s]$, where

$$p_i^s = \begin{cases} i - s_1, & i < s_1 \\ 0, & s_1 < i < s_2 \\ i - s_2, & i > s_2 \end{cases} \quad (1)$$

Here s_1 and s_2 are the starting and ending positions of the subject entity (after tokenization), and $p_i^s \in \mathbb{Z}$ is the relative distance (in tokens) to the subject entity. A position sequence relative to the object $[p_0^o, ..., p_{n+1}^o]$ can be obtained in a similar way. To incorporate the position information into the model, the position sequences are converted into position embeddings, which are then concatenated to the contextual representation $\mathcal{H}$, followed by a one-layer BiLSTM. The final hidden states in each direction of the BiLSTM are used for prediction with a one-hidden-layer MLP.

Model	P	R	F_1
Zhang et al. (2017)	65.7	64.5	65.1
Zhang et al. (2018)	69.9	63.33	66.4
Wu et al. (2019)	-	-	67.0
Alt et al. (2019)	70.1	65.0	67.4
BERT-LSTM-base	**73.3**	63.10	67.8
Zhang et al. (2018) (ensemble)	71.3	**65.4**	**68.2**

Table 1: Results on the TACRED test set.

2.2 Experiments

We evaluate our model on the TAC Relation Extraction Dataset (TACRED) (Zhang et al., 2017), a standard benchmark dataset for relation extraction. In our experiments, the hidden sizes of the LSTM and MLP are 768 and 300, respectively, and the position embedding size is 20. The learning rate is 5×10^{-5}. The BERT base-cased model is used in our experiments. Embeddings for the masks (e.g., SUBJ-LOC) are randomly initialized and fine-tuned during the training process, as well as the position embeddings.

Results on the TACRED test set are shown in Table 1. Our model outperforms the works of Zhang et al. (2018) and Wu et al. (2019), which use GCNs (Kipf and Welling, 2016) and variants to encode syntactic tree information as external features. Alt et al. (2019) leverage the pretrained language model GPT (Radford et al., 2018) and achieves better recall than our system. In terms of F_1, our system obtains the best known score among *individual* models, but our score is still below that of the interpolation model of Zhang et al. (2018) because of lower recall.

3 BERT for Semantic Role Labeling

3.1 Model

The standard formulation of semantic role labeling decomposes into four subtasks: predicate detection, predicate sense disambiguation, argument identification, and argument classification. There are two representations for argument annotation: span-based and dependency-based. Semantic banks such as PropBank usually represent arguments as syntactic constituents (spans), whereas the CoNLL 2008 and 2009 shared tasks propose dependency-based SRL, where the goal is to identify the syntactic heads of arguments rather than the entire span. Here, we follow Li et al. (2019) to unify these two annotation schemes into one framework, without any declarative constraints for

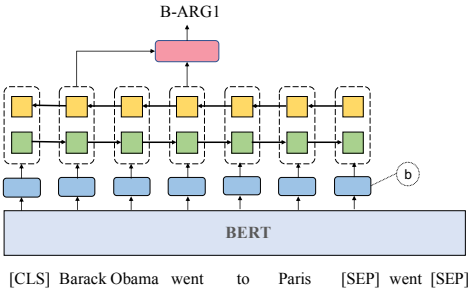

Figure 2: Architecture of our predicate identification and classification model, at the point where the model is making a prediction for the token "Barack". **(b)** denotes the concatenation of BERT contextual embedding and predicate indicator embedding. The final prediction is based on the concatenation of the hidden state of the predicate ("went") and the hidden state of the current token, fed through an MLP.

decoding. For several SRL benchmarks, such as CoNLL 2005, 2009, and 2012, the predicate is given during both training and testing. Thus, in this paper, we only discuss predicate disambiguation and argument identification and classification.

Predicate sense disambiguation. The predicate disambiguation task is to identify the correct meaning of a predicate in a given context. As an example, for the sentence "Barack Obama went to Paris", the predicate *went* has sense "motion" and has sense label *01*.

We formulate this task as sequence labeling. The input sentence is fed into the WordPiece tokenizer, which splits some words into sub-tokens. The predicate token is tagged with the sense label. Following the original BERT paper, two labels are used for the remaining tokens: 'O' for the first (sub-)token of any word and 'X' for any remaining fragments. We feed the sequences into the BERT encoder to obtain the contextual representation $\mathcal{H}$. A "predicate indicator" embedding is then concatenated to the contextual representation to distinguish the predicate tokens from non-predicate ones. The final prediction is made using a one-hidden-layer MLP over the label set.

Argument identification and classification. This task is to detect the argument spans or argument syntactic heads and assign them the correct semantic role labels. In the above example, "Barack Obama" is the ARG1 of the predicate *went*, meaning the entity in motion.

Formally, our task is to predict a sequence $\boldsymbol{z}$ given a sentence–predicate pair $(\mathcal{X}, v)$ as input,

Model	Dev	Test	Brown
Shi and Zhang (2017)	-	93.43	82.36
Roth and Lapata (2016)	94.77	95.47	-
He et al. (2018b)	95.01	95.58	-
BERT-base	**96.32**	**96.88**	**90.63**

Table 2: Predicate disambiguation accuracy on the CoNLL 2009 dataset.

Model	Dev	Test	Brown
Marcheggiani and Titov (2017)	83.3	-	-
He et al. (2018b)	84.2	-	-
Shi and Zhang (2017)	85.6	87.1	77.4
BERT-LSTM-base	88.7	89.8	82.7
BERT-LSTM-large	**89.3**	**90.3**	**83.5**

Table 3: Comparison of F_1 scores for argument identification and classification on the CoNLL 2009 dataset, excluding predicate sense disambiguation.

where the label set draws from the cross of the standard BIO tagging scheme and the arguments of the predicate (e.g., B-ARG1). The model architecture is illustrated in Figure 2, at the point in the inference process where it is outputting a tag for the token "Barack". In order to encode the sentence in a predicate-aware manner, we design the input as [[CLS] sentence [SEP] predicate [SEP]], allowing the representation of the predicate to interact with the entire sentence via appropriate attention mechanisms. The input sequence as described above is fed into the BERT encoder. The contextual representation of the sentence ([CLS] sentence [SEP]) from BERT is then concatenated to predicate indicator embeddings, followed by a one-layer BiLSTM to obtain hidden states $\mathcal{G} = [g_1, g_2, ..., g_n]$. For the final prediction on each token g_i, the hidden state of predicate g_p is concatenated to the hidden state of the token g_i, and then fed into a one-hidden-layer MLP classifier over the label set.

3.2 Experimental Setup

We conduct experiments on two SRL tasks: span-based and dependency-based. For span-based SRL, the CoNLL 2005 (Carreras and Màrquez, 2004) and 2012 (Pradhan et al., 2013) datasets are used. For dependency-based SRL, the CoNLL 2009 (Hajič et al., 2009) dataset is used. We follow standard splits for the training, development, and test sets.

In our experiments, the hidden sizes of the LSTM and MLP are 768 and 300, respectively,

	Model	CoNLL 09 (In-domain)			Out-of-domain (Brown)		
		P	R	F_1	P	R	F_1
Single	He et al. (2018b)	89.7	89.3	89.5	81.9	76.9	79.3
	Li et al. (2018)	90.3	89.3	89.8	80.6	79.0	79.8
	Li et al. (2019)	89.6	91.2	90.4	81.7	81.4	81.5
	BERT-LSTM-base	92.1	91.9	92.0	85.6	84.7	85.1
	BERT-LSTM-large	**92.4**	**92.3**	**92.4**	**85.7**	**85.8**	**85.7**
Ensemble	Roth and Lapata (2016)	90.3	85.7	87.9	79.7	73.6	76.5
	Marcheggiani and Titov (2017)	90.5	87.7	89.1	80.8	77.1	78.9

Table 4: Performance comparison on dependency-based SRL.

Model	CoNLL 05 (In-domain)			Out-of-domain (Brown)			CoNLL 12 (In-domain)		
	P	R	F_1	P	R	F_1	P	R	F_1
Strubell et al. (2018)	86.0	86.0	86.0	76.7	76.4	76.5	-	-	-
He et al. (2018a)	-	-	87.4	-	-	80.4	-	-	85.5
Ouchi et al. (2018)	88.2	87.0	87.6	79.9	77.5	78.7	87.1	85.3	86.2
Li et al. (2019)	87.9	87.5	87.7	80.6	80.4	80.5	85.7	86.3	86.0
BERT-LSTM-base	87.8	88.4	88.1	80.7	81.2	80.9	85.7	86.7	86.2
BERT-LSTM-large	88.6	**89.0**	**88.8**	**81.9**	**82.1**	**82.0**	85.9	**87.0**	86.5
Ouchi et al. (2018) (ensemble)	**89.2**	87.9	88.5	81.0	78.4	79.6	**88.5**	85.5	**87.0**

Table 5: Performance comparison on span-based SRL.

and the predicate indicator embedding size is 10. The learning rate is 5×10^{-5}. BERT base-cased and large-cased models are used in our experiments. The position embeddings are randomly initialized and fine-tuned during the training process.

3.3 Dependency-Based SRL Results

Predicate sense disambiguation. The predicate sense disambiguation subtask applies only to the CoNLL 2009 benchmark. In this line of research on dependency-based SRL, previous papers seldom report the accuracy of predicate disambiguation separately (results are often mixed with argument identification and classification), causing difficulty in determining the source of gains. Here, we report predicate disambiguation accuracy in Table 2 for the development set, test set, and the out-of-domain test set (Brown). The state-of-the-art model (He et al., 2018b) is based on a Bi-LSTM and linguistic features such as POS tag embeddings and lemma embeddings. Instead of using linguistic features, our simple MLP model achieves better accuracy with the help of powerful contextual embeddings. These predicate sense disambiguation results are used in the dependency-based SRL end-to-end evaluation.

Argument identification and classification. We provide SRL performance excluding predicate sense disambiguation to validate the source of improvements: results are shown in Table 3. Figures from some systems are missing because they only report end-to-end results.

Our end-to-end results are shown in Table 4. We see that the BERT-LSTM-large model (using the predicate sense disambiguation results from above) yields large F_1 score improvements over the existing state of the art (Li et al., 2019), and beats existing ensemble models as well. This is achieved without using any linguistic features and declarative decoding constraints.

3.4 Span-Based SRL Results

Our span-based SRL results are shown in Table 5. We see that the BERT-LSTM-large model achieves the state-of-the-art F_1 score among single models and outperforms the Ouchi et al. (2018) ensemble model on the CoNLL 2005 in-domain and out-of-domain tests. However, it falls short on the CoNLL 2012 benchmark because the model of Ouchi et al. (2018) obtains very high precision. They are able to achieve this with a more complex decoding layer, with human-designed constraints such as the "Overlap Constraint" and "Number Constraint".

4 Conclusions

Based on this preliminary study, we show that BERT can be adapted to relation extraction and

semantic role labeling without syntactic features and human-designed constraints. While we concede that our model is quite simple, we argue this is a feature, as the power of BERT is able to simplify neural architectures tailored to specific tasks. Nevertheless, these results provide strong baselines and foundations for future research. Many natural follow-up questions emerge: Can syntactic features be re-introduced to further improve results? Can multitask learning be used to simultaneously benefit relation extraction and semantic role labeling? We are actively working on answering these and additional questions.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Improving relation extraction by pre-trained language representations. In *AKBC*.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*.

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

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018a. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018b. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 411–420.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21.

Peng Shi and Yue Zhang. 2017. Joint bi-affine parsing and semantic role labeling. In *Proceedings of the 2017 International Conference on Asian Language Processing (IALP)*, pages 338–341.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038.

Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. 2019. Simplifying graph convolutional networks. *arXiv:1902.07153*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.