Multi²OIE: Multilingual Open Information Extraction based on Multi-Head Attention with BERT

Youngbin Ro  Yukyung Lee  Pilsung Kang†
Korea University, Seoul, Republic of Korea
{youngbin_ro, yukyung.lee, pilsung.kang}@korea.ac.kr

Abstract

In this paper, we propose Multi²OIE, which performs open information extraction (open IE) by combining BERT (Devlin et al., 2019) with multi-head attention blocks (Vaswani et al., 2017). Our model is a sequence-labeling system with an efficient and effective argument extraction method. We use a query, key, and value setting inspired by the Multi-modal Transformer (Tsai et al., 2019) to replace the previously used bidirectional long short-term memory architecture with multi-head attention. Multi²OIE outperforms existing sequence-labeling systems with high computational efficiency on two benchmark evaluation datasets, Re-OIE2016 and CaRB. Additionally, we apply the proposed method to multilingual open IE using multilingual BERT. Experimental results on new benchmark datasets introduced for two languages (Spanish and Portuguese) demonstrate that our model outperforms other multilingual systems without training data for the target languages.

1 Introduction

Open information extraction (Open IE) (Banko et al., 2007) aims to extract a set of arguments and their corresponding relationship phrases from natural language text. For example, an open IE system could derive the relational tuple (was elected; The Republican candidate; President) from the given sentence “The Republican candidate was elected President.” Because the extractions generated by open IE are considered as useful intermediate representations of the source text (Mausam, 2016), this method has been applied to various downstream tasks (Christensen et al., 2013; Ding et al., 2016; Khot et al., 2017; Wu et al., 2018).

Although early open IE systems were largely based on handcrafted features or fine-grained rules (Fader et al., 2011; Mausam et al., 2012; Del Corro and Gemulla, 2013), most recent open IE research has focused on deep-neural-network-based supervised learning models. Such systems are typically based on bidirectional long short-term memory (BiLSTM) and are formulated for two categories: sequence labeling (Stanovsky et al., 2018; Sarhan and Spruit, 2019; Jia and Xiang, 2019) and sequence generation (Cui et al., 2018; Sun et al., 2018; Bhutani et al., 2019). The latter enables flexible extraction; however, it is more computationally expensive than the former. Additionally, generation methods are not suitable for non-English text owing to a lack of training data because they are heavily dependent on in-language supervision (Ponti et al., 2019). Therefore, we adopted the sequence labeling method to maximize scalability by using (multilingual) BERT (Devlin et al., 2019) and multi-head attention (Vaswani et al., 2017). The main advantages of our approach can be summarized as follows:

- Our model can consider rich semantic and contextual relationships between a predicate and other individual tokens in the same text during sequence labeling by adopting a multi-head at-
MultiHead Attention for Open IE

In sequence labeling open IE systems, when extracting arguments for a specific predicate, predicate-related features are used as input variables (Stanovsky et al., 2018; Zhan and Zhao, 2019; Jia and Xiang, 2019). We analyzed this extraction process from the perspective of multimodal learning (Mangai et al., 2010; Ngiam et al., 2011; Baltrušaitis et al., 2019), which defines an entire sequence and the corresponding predicate information as a modality. The most frequently used method for open IE is simple concatenation (Figure 1, left), which can be interpreted as an early fusion approach. Simple concatenation has low computational complexity, but requires intensive feature engineering. It is also highly reliant on the choice of a classifier (Ergun et al., 2016; Liu et al., 2018).

Instead, we propose the use of a multi-modality mechanism (Tsai et al., 2019) to capture the complicated relationships between predicates and other tokens. In our method, multi-head attention is computed by using target modality as a query with source modalities as key-value pairs to adapt the latent information from sources to targets. This allows our model to assign greater weights to meaningful interactions between modalities. Accordingly, Multi2OIE uses multi-head attention to reflect predicate information (source modality) throughout a sequence (target modality). We expect this module to transform a general sentence embedding into a suitable feature for extracting the arguments associated with a specific predicate.

2.2 Multilingual Open IE

Despite the increasing amount of available web text in languages other than English, most open IE approaches have focused on the English language. For non-English languages, most systems are heavily reliant on handcrafted features and rules, resulting in limited performance (Zhila and Gelbukh, 2014; de Oliveira and Claro, 2019; Wang et al., 2019; Guarasci et al., 2020). Although some studies have demonstrated the potential of multilingual open IE (Faruqui and Kumar, 2015; Gamallo and Garcia, 2015; White et al., 2016), most approaches are based on shallow patterns, resulting in low precision (Daniela et al., 2019).

Therefore, we introduce a multilingual-BERT-based open IE system. BERT provides language-agnostic embedding through its multilingual version and provides excellent zero-shot performance on many classification and labeling tasks (Pires et al., 2019; Wu and Dredze, 2019; Karthikeyan et al., 2020). In Section 5, we demonstrate that our multilingual system yields acceptable performance when it is trained using only an English dataset.
3 Proposed Method

Multi²OIE extracts relational tuples from a given sentence in two steps. The first step is to find all predicates in the sentence. The second step is to extract the arguments associated with each identified predicate. The architecture of the proposed model is presented in Figure 2.

3.1 Task Formulation

Let $S = (w_1, w_2, ..., w_l)$ be an input sentence, where $w_i$ is the i-th token and $l$ is the sequence length. The objective of the proposed model $f$ is to find a set of tags $T = (t_1, t_2, ..., t_l)$, where each element of $T$ indicates one of the “beginning, inside, outside” (BIO) tags (Ramshaw and Marcus, 1995). However, unlike the method proposed in Stanovsky et al. (2018), which uses a predicate head as an input and predicts all tags simultaneously, we first predict a predicate tagset $T_{\text{pred}} = (t_{1}^{p}, t_{2}^{p}, ..., t_{l}^{p})$ using a predicate model $f_{\text{pred}}$. An argument tagset $T_{\text{arg}} = (t_{1}^{a}, t_{2}^{a}, ..., t_{l}^{a})$ is predicted using $f_{\text{arg}}$ based on $S$ and $T_{\text{pred}}$. Therefore, our model maximizes the following log-likelihood formulation:

$$\sum_{i=1}^{l} \left( \log p(t_{i}^{p} \mid S; \theta_{\text{pred}}) + \log p(t_{i}^{a} \mid \hat{T}_{\text{pred}}; S; \theta_{\text{pred}}; \theta_{\text{arg}}) \right),$$

where $\theta_{\text{pred}}$ and $\theta_{\text{arg}}$ are the trainable parameters of $f_{\text{pred}}$ and $f_{\text{arg}}$, respectively. In this formulation, $f_{\text{pred}}$ contributes to extracting not only the predicates, but also the arguments. The loss and gradients derived from argument extraction are also propagated to $\theta_{\text{pred}}$ and $\theta_{\text{arg}}$.

Additionally, we treat open IE as an $n$-ary extraction task and consider BIO tags for arguments up to ARG3. We refer readers to Stanovsky et al. (2018) for a more detailed explanation of the BIO sequence labeling policy.

3.2 Predicate Extraction

We assume that a given sentence $S$ is tokenized by SentencePiece (Kudo and Richardson, 2018). BERT embeds and encodes $S$ through multiple layers. The final hidden states are defined as $H \in \mathbb{R}^{l \times d}$, where $d$ is the hidden state size of BERT. $H$ is then fed into a feed-forward network and a softmax layer to calculate the probability that each token is classified into each predicate tag. The predicted tagset $\hat{T}_{\text{pred}}$ is obtained by applying the argmax operation to the softmax outputs. Finally, the loss for predicate extraction, denoted $L_{\text{pred}}$, is calculated as per-token cross-entropy loss.

3.3 Argument Extraction

A sentence contains one or more predicates. The argument extraction method described in this section
We refer to the true tagset followed by a position-wise feed-forward layer, as each of which has a multi-head attention layer followed by an argument extractor consists of $N$ multi-head attention blocks, each of which has a multi-head attention layer followed by a position-wise feed-forward layer, as shown in Figure 3.

The attention layer is the same as the encoder-decoder attention layer in the original transformer (Vaswani et al., 2017). It first transforms $X_q$, $X_k$, and $X_v$ into $Q = X_q W_q$, $K = X_k W_k$, and $V = X_v W_v$, respectively, where $W_q$, $W_k$, and $W_v$ are weight matrices with dimensions of $(d_{mh} \times d_{h})$.

Following transformation, the computation of attention is performed for each head as follows:

$$Z_h = \text{Softmax} \left( \frac{Q_h K_h^T}{\sqrt{d_h}} \right) V_h, \quad (2)$$

Each head is indexed by $h$ and has dimensions of $d_h = \frac{d_{mh}}{n_h}$, where $n_h$ denotes the number of heads. The attention outputs for each head are then concatenated and linearly transformed. In addition, we apply residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) based on the results of prior works on transformers.

The position-wise feed-forward layer consists of two linear transformations surrounding a ReLU activation function. Residual connections and layer normalization are also applied in this layer. Finally, the output of the final multi-head attention block is fed into the argument classifier. The process for obtaining a predicted argument tagset $\hat{T}_{arg}$ and corresponding argument loss $L_{arg}$ is the same as that described in Section 3.2. The final loss for parameter updating is the summation of $L_{pred}$ and $L_{arg}$.

### 3.4 Confidence Score

In open IE, confidence scores can help control the precision-recall tradeoff of a system. MultiOIE provides a confidence score for every extraction by adding the predicate score and all argument scores, as suggested in Zhan and Zhao (2019). The score of the predicate and each argument is obtained from the probability value of the $\text{Beginning}$ tag.

$$CS = p(\text{P-B}) + \sum_{i=0}^{3} p(\text{A}_{i}\text{-B}), \quad (3)$$

where the probability values are given by the softmax layer in each extraction step.

### 4 Experiments

#### 4.1 Experimental Setup

**Datasets** For fair comparisons with other systems, we trained our model using the same dataset...
Table 1: Numbers of sentences and tuples in each dataset used in this study.

| Split | Dataset   | # Sents. | # Tuples |
|-------|-----------|----------|----------|
| Train | OpenIE4   | 1,109,411| 2,175,294|
|       | OIE2016   | 582      | 1,671    |
|       | CaRB-dev  | 641      | 2,548    |
| Test  | Re-OIE2016| 595      | 1,508    |
|       | CaRB-test | 641      | 2,715    |

Table 2: Baseline models with difference settings.

| Method                  | f_pred          | f_arg          |
|-------------------------|-----------------|----------------|
| BIO                     | BIO tagging     | LSTM           |
| BIO+MH                  | BIO tagging     | LSTM           |
| SpanOIE                 | Span selection  | LSTM           |
| SpanOIE+MH              | Span selection  | LSTM           |
| BERT+BiLSTM             | BIO tagging     | BERT           |
| Multi²OIE               | BIO tagging     | BERT           |

4.2 Baselines

As baseline models, we selected RnnOIE (Stanovsky et al., 2018), SpanOIE (Zhan and Zhao, 2019), and a few custom systems to evaluate the validity of the multi-head attention blocks (MH). Although these are all sequence-labeling systems, note that SpanOIE uses the span selection method rather than BIO tagging. Table 2 presents a summary of the main baselines used in this study. We also report the results of the following systems developed prior to the use of neural networks: Stanford (Angeli et al., 2015), OLLIE (Mausam et al., 2012), PROPS (Stanovsky et al., 2016), ClausIE (Del Corro and Gemulla, 2013), and OpenIE4. For these systems, the results were from previous studies (Zhan and Zhao, 2019; Bhardwaj et al., 2019).

4.3 Results

The performance results for each system on the Re-OIE2016 and CaRB test data are presented in Table 3. The precision-recall curves are presented in Figure 4. We also present extraction examples from Multi²OIE and SpanOIE in Table 4.

Overall performance

Our model outperforms the other systems on all datasets and metrics. Our model yields average improvements of approximately 6.9%p and 2.9%p in terms of F1 for the Re-OIE2016 and CaRB datasets, respectively, compared to the state-of-the-art system (SpanOIE).

Similar to previous studies (Stanovsky et al., 2018; Zhan and Zhao, 2019), the excellent performance of Multi²OIE is attributed to improved recall. As shown in Table 3, our method achieves the highest recall rate on both datasets. The examples in Table 4 also demonstrate that our model can extract more tuples from the same sentence. An additional tuple (debut; the newly solvent airline; its new image) is found by Multi²OIE, but not by SpanOIE. Additionally, Multi²OIE extracts the place information “At a ... hangar” for the first
Figure 4: Precision-recall curves for each open IE system on two testing datasets.

|                  | Re-OIE2016 |                  | CaRB     |
|------------------|------------|------------------|----------|
|                  | AUC        | F1               | PREC.    | REC.    |
| Stanford         | 11.5       | 16.7             | -        | -       |
| OLLIE            | 31.3       | 49.5             | -        | -       |
| PropS            | 43.3       | 64.2             | -        | -       |
| ClausIE          | 46.4       | 64.2             | -        | -       |
| OpenIE4          | 50.9       | 68.3             | -        | -       |
| RnnOIE           | 68.3       | 78.7             | 84.2     | 73.9    |
| BIO              | 71.9       | 80.3             | 84.1     | 76.8    |
| BIO+MH           | 71.3       | 81.5             | 87.0     | 76.6    |
| SpanOIE          | 65.8       | 77.0             | 79.7     | 74.5    |
| SpanOIE+MH       | 68.0       | 78.8             | 83.1     | 74.9    |
| BERT+BiLSTM      | 72.1       | 81.3             | 86.0     | 77.0    |
| Multi\(^2\)OIE (ours) | 74.6 | 83.9             | 86.9     | 81.0    |

|                  | AUC        | F1               | PREC.    | REC.    |
| Stanford         | 13.4       | 23.0             | -        | -       |
| OLLIE            | 22.4       | 41.1             | -        | -       |
| PropS            | 12.6       | 31.9             | -        | -       |
| ClausIE          | 22.4       | 44.9             | -        | -       |
| OpenIE4          | 27.2       | 48.8             | -        | -       |
| RnnOIE           | 26.8       | 46.7             | 55.6     | 40.2    |
| BIO              | 27.7       | 46.6             | 55.1     | 40.4    |
| BIO+MH           | 27.3       | 47.5             | 57.2     | 40.7    |
| SpanOIE          | 30.0       | 49.4             | 60.9     | 41.6    |
| SpanOIE+MH       | 30.2       | 50.0             | 62.2     | 41.8    |
| BERT+BiLSTM      | 30.6       | 50.6             | 61.3     | 43.1    |
| Multi\(^2\)OIE (ours) | 32.6 | 52.3             | 60.9     | 45.8    |

Table 3: Performance of Multi\(^2\)OIE and baseline systems on the Re-OIE2016 and CaRB datasets.

tuple, which is omitted by SpanOIE.

**Effects of multi-head attention** We compared three pairs of methods to determine the validity of multi-head attention blocks: (BIO and BIO+MH), (SpanOIE and SpanOIE+MH), and (BERT+BiLSTM and Multi\(^2\)OIE). As a result, except for BIO+MH yielding a lower AUC than BIO, the models with multi-head attention achieve higher performance than the BiLSTM-based models. This performance improvement is consistent, regardless of the choice of classification method (BIO tagging and span selection). These results suggest that the use of multi-head attention is superior to simple concatenation in terms of utilizing predicate information.

Additionally, the performance improvement from using MH is greater with BERT than with BiLSTM. The average performance improvements from BIO to BIO+MH are -0.5%p (AUC) and 1.1%p (F1), whereas the improvements from BERT+BiLSTM to Multi\(^2\)OIE are 2.3%p (AUC) and 2.2%p (F1). This indicates that Multi\(^2\)OIE has a model architecture that can create synergies between the predicate and argument extractors.

**Computational cost** We measured the training and inference times of each system to evaluate computational efficiency. As an additional baseline model, we considered a recently published sequence generation system called IMoJIE (Kolluru et al., 2020). It achieved state-of-the-art per-
At a presentation in the Toronto Pearson International Airport hangar, Celine Dion helped the newly solvent airline debut its new image.

Multi\textsuperscript{2}OIE (helped; Celine Dion; the newly solvent airline debut its new image)

Table 6: Comparison between English (EN) and Multilingual (MT) versions of our model on CaRB dataset.

|    | AUC | F1  | PREC. | REC. |
|----|-----|-----|-------|------|
| EN version | 32.6 | 52.3 | 60.9  | 45.8 |
| MT version  | 31.5 | 51.9 | 59.5  | 45.9 |

Table 5: Training and inference times of each system.

|                 | Training | Inference | Sec./Sent. |
|-----------------|----------|-----------|------------|
| BERT+BiLSTM     | 4.5h     | 21.5s     | 0.03s      |
| SpanOIE         | 10.2h    | 33.8s     | 0.05s      |
| IMoJIE          | 7.7h     | 212.2s    | 0.33s      |
| Multi\textsuperscript{2}OIE | 4.6h | 15.5s | 0.02s |

Performance on the CaRB dataset using sequential decoding of tuples conditioned on previous extractions. For calculating inference times, we selected 641 sentences from the CaRB testing dataset and executed the models on a single TITAN RTX GPU.

Table 5 reveals that Multi\textsuperscript{2}OIE has much greater efficiency than IMoJIE. Our model only requires 15.5 s to process the 641 sentences, whereas IMoJIE requires more than 3 min, which is a difference of approximately 14 times. This bottleneck of IMoJIE could be a drawback for downstream tasks, such as knowledge base construction, which must work with large amounts of text. Considering that the performance difference between the two models is only approximately 1%p, it may be reasonable to use Multi\textsuperscript{2}OIE to process large-scale corpora. Multi\textsuperscript{2}OIE also exhibits competitive computational costs compared to the other sequence-labeling systems. Our model has similar training times compared to BERT+BiLSTM, but is faster for inference. This demonstrates that MH has a positive effect on both efficiency and performance. In the case of SpanOIE, its span selection method creates bottlenecks for both training and inference.

### 5 Multilingual Performance

As mentioned in Section 2.2, we trained a multilingual version of Multi\textsuperscript{2}OIE using multilingual BERT and the same training dataset as the English version. We assumed that data for non-English languages were not available and tested the model’s zero-shot performance. Evaluations were conducted using a dataset generated based on the Re-OIE2016 dataset.

#### 5.1 Experimental setup

**Datasets** Considering the availability of baseline systems, we selected Spanish and Portuguese as the evaluation dataset languages. First, all sentences, predicates, and arguments from the Re-OIE2016 dataset were translated into the target languages using Google\textsuperscript{5}. Following translation, we manually re-annotated all tuples of the target languages based on the English annotation of Re-OIE2016.

**Evaluation metrics** Because the baseline systems are binary extractors and do not provide confidence scores, we report binary extraction performance without AUC values. Additionally, although the introduced dataset was generated based on the Re-OIE2016, each system was tested using CaRB’s evaluation code for more rigorous evaluation.

**Baselines** Our baseline models were two rule-based multilingual systems: ArgOE (Gamallo and Garcia, 2015) and PredPatt (White et al., 2016). The former takes dependency parses in the CoNLL-X format as inputs. Similarly, the latter uses language-agnostic patterns of UD structures\textsuperscript{6}.

#### 5.2 Results

**Comparison to the English model** Prior to comparing the multilingual systems, we evaluated whether Multi\textsuperscript{2}OIE’s multilingual version exhibited a satisfactory performance for English compared to the English-only version. Table 6 lists the

---

\textsuperscript{5}https://cloud.google.com/translate/

\textsuperscript{6}https://universaldependencies.org/
When the explosion tore through the hut, Stauffenberg was convinced that no one in the room could have survived.

| Language | English | Spanish | Portuguese |
|----------|---------|---------|------------|
| Sentence | (tore; the explosion; through the hut) | (desgarró; la explosión; a través de la cabaña) | (rasgou; a explosão; através da cabana) |
|          | (was convinced; Stauffenberg; that no one in the room could have survived) | (estaba convencido; Stauffenberg; de que nadie en la habitación podría haber sobrevivido) | (estava convencido; Stauffenberg; de que ninguém na sala poderia ter sobrevivido) |
|          | (could have survived; no one in the room) | (podría haber sobrevivido; nadie en la habitación) | (poderia ter sobrevivido; ninguém na sala) |

Table 7: Extraction examples from Multi^2OIE for each language.

| Lang. | System | F1 | PREC. | REC. |
|-------|--------|----|-------|------|
| EN    | ArgOE  | 43.4 | 56.6  | 35.2 |
|       | PredPatt | 53.1 | 53.9  | 52.3 |
|       | Multi^2OIE | **69.3** | **66.9** | **71.7** |
| ES    | ArgOE  | 39.4 | 48.0  | 33.4 |
|       | PredPatt | 44.3 | 44.8  | 43.8 |
|       | Multi^2OIE | **60.2** | **59.1** | **61.2** |
| PT    | ArgOE  | 38.3 | 46.3  | 32.7 |
|       | PredPatt | 42.9 | 43.6  | 42.3 |
|       | Multi^2OIE | **59.1** | **56.1** | **62.5** |

Table 8: Binary extraction performance without confidence scores on the multilingual Re-OIE2016 dataset.

The performance metrics for the English and multilingual versions of our model on the CaRB dataset. The performance of the English version was copied from Table 3. Although the multilingual version yields lower performance for both metrics compared to the English version, the F1 score is comparable and the recall is higher. Furthermore, the multilingual version still outperforms the other sequence-labeling systems, indicating that multilingual BERT can successfully construct a Multi^2OIE model with favorable performance.

### Multilingual performance

Table 8 lists the performance metrics for each system for the multilingual dataset. Table 7 contains an example of Multi^2OIE’s extraction results for each language. One can see that Multi^2OIE outperforms the other systems on all languages. Similar to the results in Section 4.3, the superiority of our multilingual model is attributed to its high recall. Multi^2OIE yields the highest recall for all languages by approximately 20%. In contrast, ArgOE has relatively high precision, but low recall negatively impacts its F1 score. PredPatt provides the best balance of precision and recall, but the overall performance is lower than that of our model.

The performance differences between languages are similar for all models. All models exhibit the best performance for English, followed by Spanish and Portuguese. Multi^2OIE also exhibits performance degradation for non-English languages. However, considering that our model was never trained to perform open IE tasks on Spanish or Portuguese, its performance is remarkable. For some non-English sentences, our model extracts the same results as those extracted in the English extraction result, as shown in Table 7. This result agrees with the results of previous studies (Pires et al., 2019; Wu and Dredze, 2019; Karthikeyan et al., 2020), which have demonstrated the excellent cross-lingual abilities of multilingual BERT. Based on these results, we expect that Multi^2OIE will also work well on languages other than those considered in this study.

### 6 Conclusion

In this paper, we propose Multi^2OIE, which exploits BERT and multi-head attention for the open IE task. Multi-head attention has the advantage of fusing sentence and predicate features, which adequately reflect predicate information throughout a sentence. Our model achieved the best performance among sequence labeling models. Multi^2OIE also exhibited superior computational efficiency with competitive performance compared to the state-of-the-art sequence generation systems. Additionally, a Multi^2OIE model trained using multilingual BERT, outperformed the baseline models without training on any non-English languages.

However, some types of extractions, such as
nominal relations, conjunctions in arguments, and contextual information, are not considered in Multi²OIE. Future work could investigate how to apply Multi²OIE to these cases. For multilingual open IE, performance evaluations and further study on non-alphabetic languages that were not considered in this study can be conducted.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 344–354, Beijing, China. Association for Computational Linguistics.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. ArXiv, abs/1607.06450.

T. Baltrušaitis, C. Ahuja, and L. Morency. 2019. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):423–443.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. In Proceedings of the 2019 Conference onEmpirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Nikita Bhutani, Yoshihiko Suhara, Wang-Chiew Tan, Alon Halevy, and H. V. Jagadish. 2019. Open information extraction from question-answer pairs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2294–2305, Minneapolis, Minnesota. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multidocument summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1163–1173, Atlanta, Georgia. Association for Computational Linguistics.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 407–413, Melbourne, Australia. Association for Computational Linguistics.

Barreiro C. Daniela, Marlo Souza, Clarissa C. Xavier, and Leandro Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. Copyright - © 2019. This work is published under https://creativecommons.org/licenses/by/4.0/ (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated -2019-05-07.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In Proceedings of the 22nd International Conference on World Wide Web, WWW ’13, page 355–366, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2133–2142, Osaka, Japan. The COLING 2016 Organizing Committee.

Hilal Ergun, Yusuf Caglar Akyuz, Mustafa Sert, and Jianqian Liu. 2016. Early and late level fusion of deep convolutional neural networks for visual concept recognition. International Journal of Semantic Computing, 10(03):379–397.

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, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In Progress in Artifi-
Raffaele Guarasci, Emanuele Damian, Aniello Minuto, Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian. 2019. Hybrid neural tagging model for open relation extraction.

Kaliyaperumal Kolluru, Samarth Aggarwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 711–722, Cham. Springer International Publishing.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. ArXiv, abs/1706.02677.

Raffaele Guarasci, Emanuele Damian, Aniello Minuto, Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian. 2019. Hybrid neural tagging model for open relation extraction.

Kaliyaperumal Kolluru, Samarth Aggarwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 711–722, Cham. Springer International Publishing.

Kaiming He, Xiangyu Zhang, Shaqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

Shengbin Jia and Yang Xiang. 2019. Hybrid neural tagging model for open relation extraction.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam Mausam, and Soumen Chakrabarti. 2020. Cross-lingual ability of multilingual bert: An empirical study. ArXiv, abs/1912.07840.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. CoRR, abs/1704.05572.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam Mausam, and Soumen Chakrabarti. 2020. Cross-lingual ability of multilingual bert: An empirical study. ArXiv, abs/1912.07840.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. CoRR, abs/1704.05572.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam Mausam, and Soumen Chakrabarti. 2020. Cross-lingual ability of multilingual bert: An empirical study. ArXiv, abs/1912.07840.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. CoRR, abs/1704.05572.

Toku Kudo and John Richardson. 2018. Sentence-Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kuan Liu, Yanen Li, Ning Xu, and Premkumar Natarajan. 2018. Learn to combine modalities in multimodal deep learning. ArXiv, abs/1805.11730.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In ICLR.

Uthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. 2010. A survey of decision fusion and feature fusion strategies for pattern classification. Iete Technical Review, 27:293–307.

Mausam Mausam. 2016. Open information extraction systems and downstream applications. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16, page 4074–4077. AAAI Press.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11, page 689–696, Madison, WI, USA. Omnipress.

Leandro Souza de Oliveira and Daniela Barreiro Claro. 2019. Dptoie: a portuguese open information extraction system based on dependency analysis.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, page III–1310–III–1318. JMLR.org.

Telmo Pires, Eva Schlünger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019. Towards zero-shot language modeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2900–2910, Hong Kong, China. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In Third Workshop on Very Large Corpora.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Injy Sarhan and Marco R. Spruit. 2019. Contextualized word embeddings in a neural open information extraction model. In NLPDB.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2300–2305, Austin, Texas. Association for Computational Linguistics.

Gabriel Stanovsky, Jessica Fieler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. ArXiv, abs/1603.01648.
Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. **Supervised open information extraction.** 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 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018. **Logician: A unified end-to-end neural approach for open-domain information extraction.** In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM ’18*, page 556–564, New York, NY, USA. Association for Computing Machinery.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. **Multimodal transformer for unaligned multimodal language sequences.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefined, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.** In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

C. Wang, X. He, and A. Zhou. 2019. **Open relation extraction for Chinese noun phrases.** *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Aaron Steven White, Drew Reisinger, Keisuke Sagaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. **Universal Decompositional Semantics on Universal Dependencies.** In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

R. J. Williams and D. Zipser. 1989. **A learning algorithm for continually running fully recurrent neural networks.** *Neural Computation*, 1(2):270–280.

Shijie Wu and Mark Dredze. 2019. **Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Tien-Hsuan Wu, Zhiyong Wu, Ben Kao, and Pengcheng Yin. 2018. **Towards practical open knowledge base canonicalization.** In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM ’18*, page 883–892, New York, NY, USA. Association for Computing Machinery.

Junlang Zhan and Hai Zhao. 2019. **Span model for open information extraction on accurate corpus.** *arXiv: Computation and Language*.

Alisa Zhila and Alexander Gelbukh. 2014. **Open information extraction for Spanish language based on syntactic constraints.** In *Proceedings of the ACL 2014 Student Research Workshop*, pages 78–85, Baltimore, Maryland, USA. Association for Computational Linguistics.