Extensions to BERT: SpanBERT (Joshi et al.) and Cross-Lingual Language Model Pretraining (Lample and Conneau)

Owain West

University of Waterloo

February 3, 2020

Table of contents

Plan

We will discuss two modifications to the BERT pretraining setup which improve its performance.

One is a the addition of another local pretraining objective: *Span prediction*

The other is cross-lingual training, either on a collection of various languages ("unsupervised cross-lingual pretraining"), or on sentence pairs from different languages ("supervised cross-lingual pretraining")

SpanBERT

SpanBERT (Joshi, Chen, Liu, Weld, Zettlemoyer, and Levy)

In short: modifies BERT to mask contiguous spans of tokens, and adds a related pretraining objective

Contributions:

- Modifies BERT to mask contiguous spans of tokens
- Introduces a corresponding pretraining objective which predicts tokens in the masked span solely from the tokens immediately preceding and following the span
- Shows SpanBERT outperforms BERT and other baselines when trained on the same data, and achieving new SotA results on various downstream tasks

Previous Work

SpanBERT directly builds on BERT

It uses the some lessons from other BERT modifications, eg not using the NSP task

Span Masking

Given a sequence $X = \langle x_1 ... x_n \rangle$ of tokens, a subset $Y \subseteq X$ (with $|Y|/|X| \leq 0.15$) is picked by randomly choosing contiguous spans of tokens

Span masking always begins at the a token corresponding to a new word

Span Masking

The length of a given span is chosen according to geometric distribution $P(k) = (1-p)^{k-1}p$ clipped at $k = 10$ and with $p = 0.2$. Avg. span length was $l = 3.8$ words

k was measured in terms of whole words (not tokens) masked

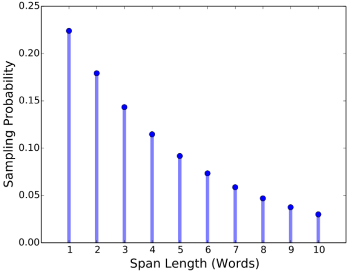


Figure 2: We sample random span lengths from a geometric distribution $\ell \sim Geo(p = 0.2)$ clipped at $\ell_{max} = 10$.

Span Boundary Objective

Given a masked subsequence $\langle x_s \ldots x_e \rangle$, every internal token is represented as a function of the tokens immediately preceding and following the span, as well as a positional embedding

$$\mathrm{y}_i = f(\mathrm{x}_{s-1}, \mathrm{x}_{e+1}, \mathrm{p}_i)$$

Here, f is $LayerNorm(GeLU(W_2 \cdot h))$ where $h = LayerNorm(GeLU(W_1 \cdot [x_{s-1}; x_{e+1}; p_i]))$ and W_1, W_2 are trainable weights

Cross-entropy loss is used for the SBO (as for the masked-LM loss) and then added to the total loss

Pretraining

They reimplemented BERT as their baseline. The same model configuration as BERT-xlarge was used, and trained on the same two corpi (BooksCorpus and English Wikipedia)

Differences:

- Different masks were used each epoch, as opposed to selecting the masked tokens when creating pretraining data
- Unlike BERT, shorter sequences were not selected with probability 0.1
- Optimized for 2.4M steps with learning rate of 1×10^{-8}, taking 15 days on 32 V100 GPUs

Experiments

Fine-tuned the resultant models on:

- Extractive question answering
- Coreference resolution
- Relation extraction
- GLUE tasks

Baselines

Baselines were:

- Original BERT model
- Reimplementation of BERT, with NSP and sentence pairs
- BERT reimplementation, without NSP and with single sentences

Extractive QA

Tested on the SQUAD 1.1, SQUAD 2.0, and MRQA datasets
QA pair is encoded as $[CLS]p_q...p_l[SEP]q_1...q_m[SEP]$, and linear classifiers are added to predict the answer span start and end

	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Human Perf.	82.3	91.2	86.8	89.4
Google BERT	84.3	91.3	80.0	83.3
Our BERT	86.5	92.6	82.8	85.9
Our BERT-1seq	87.5	93.3	83.8	86.6
SpanBERT	**88.8**	**94.6**	**85.7**	**88.7**

Table 1: Test results on SQuAD 1.1 and SQuAD 2.0.

	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQA	(Avg)
Google BERT	68.8	77.5	81.7	78.3	79.9	77.3
Our BERT	71.0	79.0	81.8	80.5	80.5	78.6
Our BERT-1seq	71.9	80.4	84.0	80.3	81.8	79.7
SpanBERT	**73.6**	**83.6**	**84.8**	**83.0**	**82.5**	**81.5**

Table 2: Performance (F1) on the five MRQA extractive question answering tasks.

Coreference Resolution

This is the task of clustering various tokens with the same referent. Tested on the CoNLL-2012 shared task.

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Prev. SotA: (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
Our BERT	85.1	83.5	84.3	77.3	75.5	76.4	75.0	71.9	73.9	78.3
Our BERT-1seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
SpanBERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

Table 3: Performance on the OntoNotes coreference resolution benchmark. The main evaluation is the average F1 of three metrics – MUC, B^3, and $CEAF_{\phi_4}$ on the test set.

Relation Extraction

This is the task of predicting the relation between two given spans of text within a sequence. Tested on the TACRED dataset.

	P	R	F1
Curr. SotA: (Soares et al., 2019)	-	-	71.5
Google BERT	69.1	63.9	66.4
Our BERT	67.8	67.2	67.5
Our BERT-1seq	**72.4**	67.9	70.1
SpanBERT	70.8	**70.9**	**70.8**

Table 5: Test set performance on the TACRED relation extraction benchmark.

GLUE Tasks

GLUE is a standard set of language understanding benchmarks, including single-sentence tasks, similarity tasks, and inference tasks.

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	(Avg)
Google BERT	59.3	**95.2**	88.5/84.3	86.4/88.0	71.2/89.0	86.1/85.7	93.0	71.1	80.4
Our BERT	58.6	93.9	90.1/86.6	88.4/89.1	71.8/89.3	87.2/86.6	93.0	74.7	81.1
Our BERT-1seq	63.5	94.8	**91.2**/87.8	89.0/88.4	**72.1/89.5**	88.0/87.4	93.0	72.1	81.7
SpanBERT	**64.3**	94.8	90.9/**87.9**	**89.9/89.1**	71.9/**89.5**	**88.1/87.7**	**94.3**	**79.0**	**82.8**

Table 4: Test set performance metrics on GLUE tasks. MRPC: F1/accuracy, STS-B: Pearson/Spearmanr correlation, QQP: F1/accuracy, MNLI: matched/mistached accuracies. WNLI (not shown) is always set to majority class (65.1% accuracy) and included in the average.

CoLA: Corpus of Linguistic Accaptability
SST-2: Stanford Sentiment Treebank MNLI: Multi-Genre Natural Language Inference
QNLI: SQUAD as binary classification
RTE: recognizing textual entailment

Ablation Studies

Various masking schemes were tested, including:

- Subword tokens: sample individual tokens
- Whole words: sample whole words
- Named entities: 50% of the time, sample a named entity[1]; 50% of the time, sample a random word
- Noun phrases: 50% of the time, sample a noun phrase[2]; 50% of the time, sample a random word
- Random spans: as in SpanBERT

The effects of NSP, lack of NSP, and SBO are tested.

[1] Using spaCy NER
[2] Using spaCy constituency parser

Ablation Studies

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Subword Tokens	83.8	72.0	76.3	**77.7**	86.7	92.5
Whole Words	84.3	72.8	77.1	76.6	86.3	92.8
Named Entities	84.8	72.7	78.7	75.6	86.0	93.1
Noun Phrases	85.0	**73.0**	77.7	76.7	86.5	93.2
Random Spans	**85.4**	**73.0**	**78.8**	76.4	**87.0**	**93.3**

Table 6: The effect of replacing BERT's original masking scheme (Subword Tokens) with different masking schemes. Results are F1 scores for QA tasks and accuracy for MNLI and QNLI on the development sets. All the models are based on bi-sequence training with NSP.

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Span Masking (2seq) + NSP	85.4	73.0	78.8	76.4	87.0	93.3
Span Masking (1seq)	86.7	73.4	80.0	76.3	87.3	93.8
Span Masking (1seq) + SBO	**86.8**	**74.1**	**80.3**	**79.0**	**87.6**	**93.9**

Table 7: The effects of different auxiliary objectives, given MLM over random spans as the primary objective.

Discussion

Gives further credence to the idea that well-designed pretraining objectives are semantically meaningful

Would testing a variable-width context for the SBO prediction be worthwhile?

Can you think of other similar pretraining objectives which may be relevant to sentence structure?

XLM

Cross-Lingual Language Model Pretraining (Lample and Conneau)

In short: applies BERT to cross-lingual language modelling.

Contributions:

- Introduces unsupervised and supervised BERT-based cross-lingual pretraining objectives
- Shows both objectives lead to an improvement on a number of cross-lingual tasks
- Shows cross-lingual models especially help with low-resource (ie small corpus) languages
- Releases code and models

Previous Work

Unsupervised pretraining has been shown to be effective on a number of tasks, especially in connection with Transformer models

Previous work has been mostly in English (monolingual)

The authors have previously done some work on cross-lingual models, and have released a test set, XNLI (Cross-lingual Natural Language Inference corpus)

Previous Work

There has been substantial previous work on aligning text embeddings, mostly in a supervised fashion

Some recent work has reduced the need for supervised cross-lingual pretraining, showing embeddings can be aligned in an unsupervised manner

When substantial parallel data is available, supervised approaches can work well even for zero-shot translation

Models

Trained three different models: CLM (Causal Language Model), MLM (Masked Language Model), and TLM (Translation Language Model).

CLM and MLM are trained with monolingual data, whereas TLM is trained with parallel data from a multilingual corpus

Notation: we have a corpus C_i, $1 \leq i \leq N$, for each of N languages, and denote $|C_i| = n_i$.

Models: Preprocessing

Data from the multilingual corpus is tokenized by byte-pair encoding with a shared vocabulary

The BPE tokens are learned from concatenations of sentences sampled a single monolingual corpus

The monolingual corpus to select from is picked with probability q_i of a multinomial distribution with parameters $q_i = \frac{p_i^\alpha}{\sum_{1 \leq j \leq N} p_j^\alpha}, p_i = \frac{n_i}{\sum_{1 \leq k \leq N} n_k}$

Models: CLM

The *Causal Language Model* is just a standard left-to-right Transformer-based LM

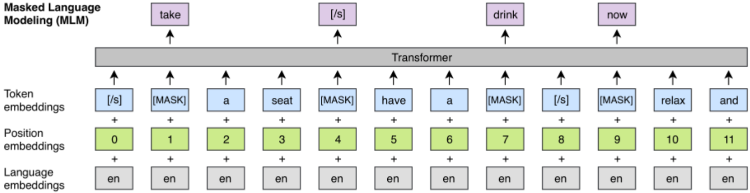
It optimizes θ so as to maximize $P(w_t|w_1...w_{t-1}, \theta)$

The first words in each batch are treated as being without context

Models: MLM

The *Masked Language Model* is trained on the standard BERT cloze (language model masking) task

Differences: uses streams of 256 tokens (not sentence pairs), and subsamples tokens to mask according to inverse frequency

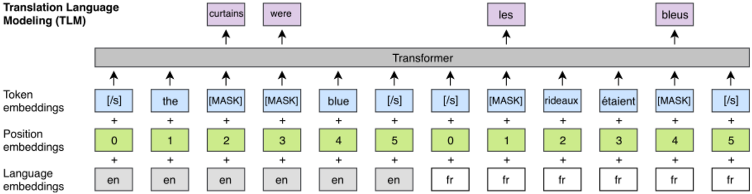


Models: TLM

The *Translation Language Model* is the only model introduced which requires parallel data

Trains on the concatenation of parallel sentences, split by a separator token

Intuitively, this allows the model to attend to the foreign context to infer a word when the monolingual context is insufficient



Training Details: Model

Uses a Transformer with:

- 1024 hidden units
- 8 heads for multi-headed-attention
- GELU
- Dropout of 0.1
- LR between $1 - 5.1 \times 10^{-4}$

Seq length of 256 and mini-batches of size 64 for CLM and MLM

Mini-batches of approximately 4000 tokens of similar-length sentences for TLM

Training Details: Data

Monolingual data was obtained by *WikiExtractor*

Parallel data was the

- MultiUN corpus for Arabic, Chinese, French, Russian, and Spanish
- IIT Bombay Corpus for Hindi
- EUBookShop Corpus for Bulgarian, German, and Greek
- OpenSubtitles 2018 corpus for Thai, Turkish, and Vietmanese
- Tanzil corpus for Swahili and Urdu
- GlobalVoices corpus for Swahili

Experiments

Fine-tuned the resultant models on:

- Cross-lingual classification
- Unsupervised translation
- Supervised translation

They also discuss the effects of cross-lingual pretraining on low-resource languages

Cross-Lingual Classification

Evaluates on the XNLI (Cross-Lingual Natural Language Inference) dataset, which contains 5000 test and 2500 dev pairs annotated with textual entailment in each of 15 languages

Adds a linear layer above the first pretrained hidden layer. Fine-tunes on English NLI dataset, tests on the other 15 languages

Results of two translation benchmarks from XNLI are also reported

Cross-Lingual Classification: XNLI

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction

TRANSLATE-TRAIN: translates english into target language at training time, then learns classifiers
TRANSLATE-TEST: target languages are translated to English, then fed into an English classifier

Results: Cross-Lingual Classification

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

Unsupervised MT

Evaluates on WMT '14 English-French, WMT '16 English-German, and WMT '16 English-Romanian

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lampe et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Unsupervised MT

Algorithm 1: Unsupervised MT

Language models: Learn language models P_s and P_t over source and target languages;

Initial translation models: Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P^{(0)}_{s \to t}$ and $P^{(0)}_{t \to s}$;

for k=1 **to** N **do**

  Back-translation: Generate source and target sentences using the current translation models, $P^{(k-1)}_{t \to s}$ and $P^{(k-1)}_{s \to t}$, factoring in language models, P_s and P_t;

  Train new translation models $P^{(k)}_{s \to t}$ and $P^{(k)}_{t \to s}$ using the generated sentences and leveraging P_s and P_t;

end

Supervised MT

Evaluates on WMT '16 Romanian-English

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro $\rightarrow$ en	28.4	31.5	35.3
ro $\leftrightarrow$ en	28.5	31.5	35.6
ro $\leftrightarrow$ en + BT	34.4	37.0	**38.5**

Table 3: **Results on supervised MT.** BLEU scores on WMT'16 Romanian-English. The previous state-of-the-art of Sennrich et al. (2016) uses both back-translation and an ensemble model. ro $\leftrightarrow$ en corresponds to models trained on both directions.

Low-Resource Languages

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

Table 4: **Results on language modeling.** Nepali perplexity when using additional data from a similar language (Hindi) or a distant one (English).

Similarity Comparison to other Cross-Lingual Embeddings

- MUSE - uses adversarial learning to align monolingual embeddings
- Concat - fastText embedding to concatenation of monolingual corpora

	Cosine sim.	L2 dist.	SemEval'17
MUSE	0.38	5.13	0.65
Concat	0.36	4.89	0.52
XLM	**0.55**	**2.64**	**0.69**

Table 5: **Unsupervised cross-lingual word embeddings** Cosine similarity and L2 distance between source words and their translations. Pearson correlation on SemEval'17 cross-lingual word similarity task of Camacho-Collados et al. (2017).

Discussion

There is useful shared information to be gained by cross-lingual language modelling

This paper shows that cross-lingual pretraining is possible in a fully unsupervised fashion, and additionally gives a new strong method for supervised cross-lingual LM pretraining

Protein LMs

Language Modelling of Proteins

Proteins are sequences of amino acids (a 20-character alphabet)

Their sequence data encodes information about their structure and function

So: treat proteins exactly like you would a natural language

Previous work has mainly used HMMs and LSTMs

Language Modelling of Proteins

There has been some work on pretraining with deep unsupervised embeddings (eg with ELMo) but none in depth that I'm aware of

- None pretrained on a large-scale dataset
- None which takes protein features into account in the LM pretraining
- None which give a thorough study of the effects of data representation choices (eg BPE vs contiguous-token vs overlapping-token embeddings) or model parameters (eg sequence length) in downstream applications

Language Modelling of Proteins: Downstream Goals

The holy grail is drug design

Protein sequencing (translation of MS/MS data to amino acid), especial for novel/highly variable proteins (eg antibodies, monoclonal and polyclonal)

Functional annotation

Structural prediction

Language Modelling of Proteins: Current Gaps

There has been some work on pretraining with deep unsupervised embeddings (eg with ELMo) but none in depth that I'm aware of

- None pretrained on a large-scale dataset
- None which takes protein features into account in the LM pretraining
- None which give a thorough study of the effects of data representation choices (eg BPE vs contiguous-token vs overlapping-token embeddings) or model parameters (eg sequence length) in downstream applications

There has been some finetuning of English BERT models for healthcare-specific text

Relevant Protein-Specific Language Features?

Local: hydrophobicity, charge, solubility

Global: protein-pair "same family" task

Others?

Project: Goals

Current status: have pretrained models for Uniprot PE1s (approx 150,000 proteins) and PE2s (approx 1.5mil) with non-overlapping tokenizations into n-grams for $n \in \{1, 2, 3, 4\}$, as well as PE1 1, 3gram with 3-way hydrophobicity classification. Each takes around 50hrs on a TPUv2

Ongoing goal: determine best protein-specific pretraining procedures on smaller data, continually testing on structural and functional classification tasks as models are available, and sequencing task once finished implementing

Ultimate goal: pretrain on largest reasonable dataset and apply to generative tasks

Project: Goals

Current limitations:

- Currently only have pretrained on a relatively small amount of pretraining data (compared to, eg, the whole Uniprot dataset of approx. 180mil proteins)
- Have only been able to pretrain with seq length=128, and have not trained models more than 1mil steps
- Have not yet been able to adequately test the effects of various learning rates
- Have not yet been able to test with overlapping or BPE tokenization

Project: Ongoing Results

Protein *family*: corresponds to structurally similar proteins with recent common ancestor (approx 5000 superfamilies in SCOPe). Protein *family*: corresponds to structurally similar proteins (approx 2000 superfamilies in SCOPe)

Finetuning the PE1-1gram model for only approx. 30 mins on a TPUv2 results in a model comparable to the current SoTA on binary classification of sequences from the same superfamily into "same family" / "different family"

Observations: 1gram-model is currently better than higher *n*-gram models. More training data (PE2 vs PE1) helps. Local hydrophobicity prediction helps, although not as much as more training data does

Project: Tasks

Ongoing task: implement graph2seq model to finetune current pretrained models for sequencing (converting graph of possible adjacent ngram subsequences to protein); collect training data

Ongoing task: train and test models with additional local and global protein-specific features

Ongoing task: port current BERT-based implementation to AlBERT for more time-efficient pretraining

Future task: implement span prediction

Future task: swap out a Reformer for the Transformer in AlBERT, and pretrain on longer protein sequences

Conclusion

Additional pretraining objectives can be useful

Further results of the protein BERT work will be presented as my class project

Appendix

Hyperparameters

Extractive QA: Learning rates chosen among $\{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$. Max sequence length set to 512. 4 epochs with batch sizes among $\{16, 32\}$

Coreference resolution: max seq length chosen among $\{128, 256, 384, 512\}$. BERT learning rate among $\{1 \times 10^{-5}, 2 \times 10^{-5}\}$ and task-specific learning rates among $\{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}\}$, 20 epochs with batch size of 1

GLUE and Relation Extraction: learning rates chosen among $\{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$. Max sequence length set to 512. 10 epochs with batch sizes among $\{16, 32\}$ (except for CoLA, with 4 epochs)

Perplexity

Perplexity measures how good a distribution is at predicting samples; a lower score indicates more accurate predictions. The perplexity of a distribution P is

$$2^{-\sum_x P(x) \lg P(x)}$$

SQUAD

SQUAD 1: 100,000 questions. SQUAD 2: adds 50,000 unanswerable

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

MRQA

A collection of a number of question-answering datasets. Used:

- NewsQA
- SearchQA
- TriviaQA
- HotpotQA
- Natural Questions

Coreference Resolution

Built off of a higher-order coreference model from *BERT for Coreference Resolution: Baselines and Analysis*

A span x is associated with possible referrent spans y, and the model is trained to predict the probability distribution of the y's given x (represented as a softmax of a scoring function of pairs (x, y))

The scoring function $s(x, y)$ is computed by taking into account the likelihood of x, the likelihood of y, and the joint probability of x and y.

Relation Extraction

This is the task of predicting the relation between two given spans of text within a sequence. Tested on the TACRED dataset.

Example	Entity Types & Label
Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said.	**Types:** PERSON/TITLE **Relation:** *per:title*
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	**Types:** PERSON/CITY **Relation:** *per:city_of_birth*
Pandit worked at the brokerage Morgan Stanley for about 11 years until 2005, when he and some Morgan Stanley colleagues quit and later founded the hedge fund Old Lane Partners.	**Types:** ORGANIZATION/PERSON **Relation:** *org:founded_by*
Baldwin declined further comment, and said JetBlue chief executive Dave Barger was unavailable.	**Types:** PERSON/TITLE **Relation:** *no_relation*

Table 1: Sampled examples from the TACRED dataset. Subject entities are highlighted in blue and object entities are highlighted in red.

GLUE Tasks

GLUE is a standard set of language understanding benchmarks, including single-sentence tasks, similarity tasks, and inference tasks.

Corpus	|Train|	|Dev|	|Test|	Task	Metric	Domain
			Single-Sentence Tasks			
CoLA	10k	1k	1.1k	acceptability	Matthews	linguistics literature
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	4k	N/A	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman	misc.
QQP	400k	N/A	391k	paraphrase	acc./F1	social QA Questions
			Inference Tasks			
MNLI	393k	20k	20k	NLI	acc. (match/mismatch)	misc.
QNLI	108k	11k	11k	QA/NLI	acc.	Wikipedia
RTE	2.7k	N/A	3k	NLI	acc.	misc.
WNLI	706	N/A	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-Benchmark, which is a regression task. MNLI has three classes while all other classification tasks are binary.

Unsupervised MT

- Language modelling is done by autoencoding noisy tokens $\mathcal{L}^{lm} = \mathbb{E}_{x\sim S}[-\log P_{s\to s}(x|C(x))] + \mathbb{E}_{y\sim T}[-\log P_{t\to t}(y|C(y))]$
- Back-translation minimizes the loss of translating purported translations x^*, y^* of source/target sentences x, y respectively back into their original x, y. Formally, $\mathcal{L}^{back} = \mathbb{E}_{y\sim T}[-\log P_{s\to t}(y|y^*)] + \mathbb{E}_{x\sim S}[-\log P_{t\to s}(x|x^*)]$

References I

- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer: SpanBERT: Improving Pre-training by Representing and Predicting Spans, 2019. http://arxiv.org/abs/1907.10529. arXiv:1907.10529
- Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925.
- Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jegou. 2018a. Word translation without parallel data. In ICLR.

References II

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In Proceed- ings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 328–339.

References III

- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based neural unsupervised machine translation. In EMNLP.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

References IV

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language- unsupervised/language understanding paper.pdf.
- Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. arXiv preprint arXiv:1611.02683.

References V

- William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. 2019. KERMIT: Generative insertion-based modeling for sequences. arXiv preprint arXiv:1906.01604.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

References VI

- Yu Stephanie Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin- Iun Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In Association for Compu- tational Linguistics (ACL), pages 1441–1451.

References VII

- Position-aware Attention and Supervised Data Improve Slot Filling Y Zhang, V Zhong, D Chen, G Angeli, CD Manning EMNLP 2017
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, 2018. http://arxiv.org/abs/1804.07461. arXiv:1804.07461