CLOZE Answer Generator

Jimmy Zhou, Javen Xu, Jingbo Yang

zhouxq, javenxu, jingboy @stanford.edu

CS224N Final Project, Department of Computer Science, Stanford University

Stanford Computer Science

Motivation

Sentence completion is critical for construction of language models. High quality cloze-style dataset with multiple sentences and multiple blanks only came into existence in 2018. An accurate blank filling model not only demonstrates capability of advanced language models, but will also help OCR tasks and hand writing recognition that benefit from longer context. We reworked existing models including ELMo and BERT then analyzed performance for various scenarios.

Dataset

The CLOTH dataset [1] is a collection of cloze tests containing 7,131 articles and 99,433 questions. These samples were collected from online resources for English examinations in China. Articles are split using 70:15:15 ratio.

People's fingernails and toenails, **(0)** ..... to a recent study, are nowadays growing more quickly. Research **(1)** ..... out at the University of North Carolina indicates that the speed at which human nails are growing has increased by **(2)** ..... to 25 per cent over the last 70 years.

The results of the study show that the **(3)** ..... human fingernail now grows about 3.5 mm a month, **(4)** ..... with just 3 mm seven decades ago. Toenail growth, **(5)** ..... only about 2 mm per month, was also up on the figure **(6)** ..... in a similar survey done 70 years ago.

1	**A** taken	**B** carried	**C** studied	**D** worked
2	**A** near	**B** just	**C** close	**D** next
3	**A** average	**B** medium	**C** common	**D** standard
4	**A** opposed	**B** measured	**C** related	**D** compared
5	**A** although	**B** despite	**C** however	**D** nevertheless
6	**A** achieved	**B** concluded	**C** arrived	**D** obtained

Figure 1. Sample cloze test

Two additional datasets were also created. One uses original CLOTH articles but options are replaced by closest neighbours using GLoVE embedding. Articles for the other dataset are made of sentences from 1BLM. Blanks are randomly selected and options are GLoVE closest neighbours.

	Options for "although"		
CLOTH	despite	however	nevertheless
GLoVE 50D	as	however	though
GLoVE 300D	fact	however	though

Table 1. Comparison of options

Approach

We first dissected and reproduced the architecture of an adapted BERT model [2], developed by authors of the CLOTH dataset. This model uses a pre-trained BERT encoder, where each word is converted to 3-length token. The decoder is a fully-connected linear layers which turns the 768-length embedding vector into a 30552 long vector, corresponding to the vocabulary space.

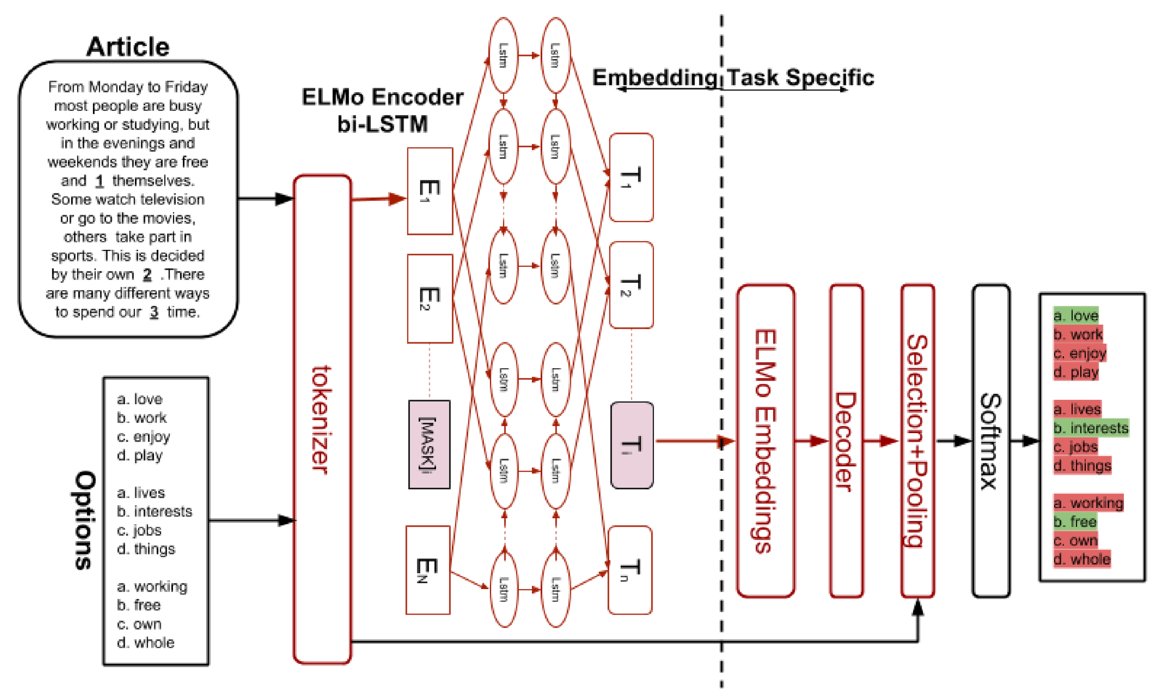

Figure 2.ELMo model architecture

Both embedding layer and decoder can be replaced by custom architectures with adequate tokenizer and result interpreter.

The BERT tokenizer was converted to generate character-level encoding for ELMo [3]. Specifically, the character embedding is context aware and maps to a 50 dimensional space. At the encoder level, ELMo outputs a 1024-length embedding vector. At the decoder level, we again used linear layers to turn the embedding into 28k long decoded vectors representing the ELMo vocabulary space.

A bidirectional TCN model was created as an attempt to replace LSTM structure used for ELMo. All other components of the architecture are kept the same. This model was trained using the same method as BERT.

Results and Discussion

Model	Training Accuracy	Testing Accuracy
Custom TCN	95%	39%
LSTM	82%	48%
Attentive Reader	95%	53%
BERT Base	96%	83%
BERT Large	95%	86%
Adapted ELMo	100%	53%
Sentence Classification	N/A	58%

Table 2. Model performance

It comes with no surprise that BERT models are the highest performing because BERT was trained to recover masked tokens. The ELMo model is on-par with previous baseline, while identifying answer through sentence classification exceeded baselines by 5%. Although additional datasets were used, overfitting was still present for all non-BERT models, suggesting that alternative training methods should be considered.

Learning rate scheduling is also important. As shown in Figure 3, jumps in accuracy were observed when learning rate stepped down. Similar effects were also observed for training TCN and BERT models.

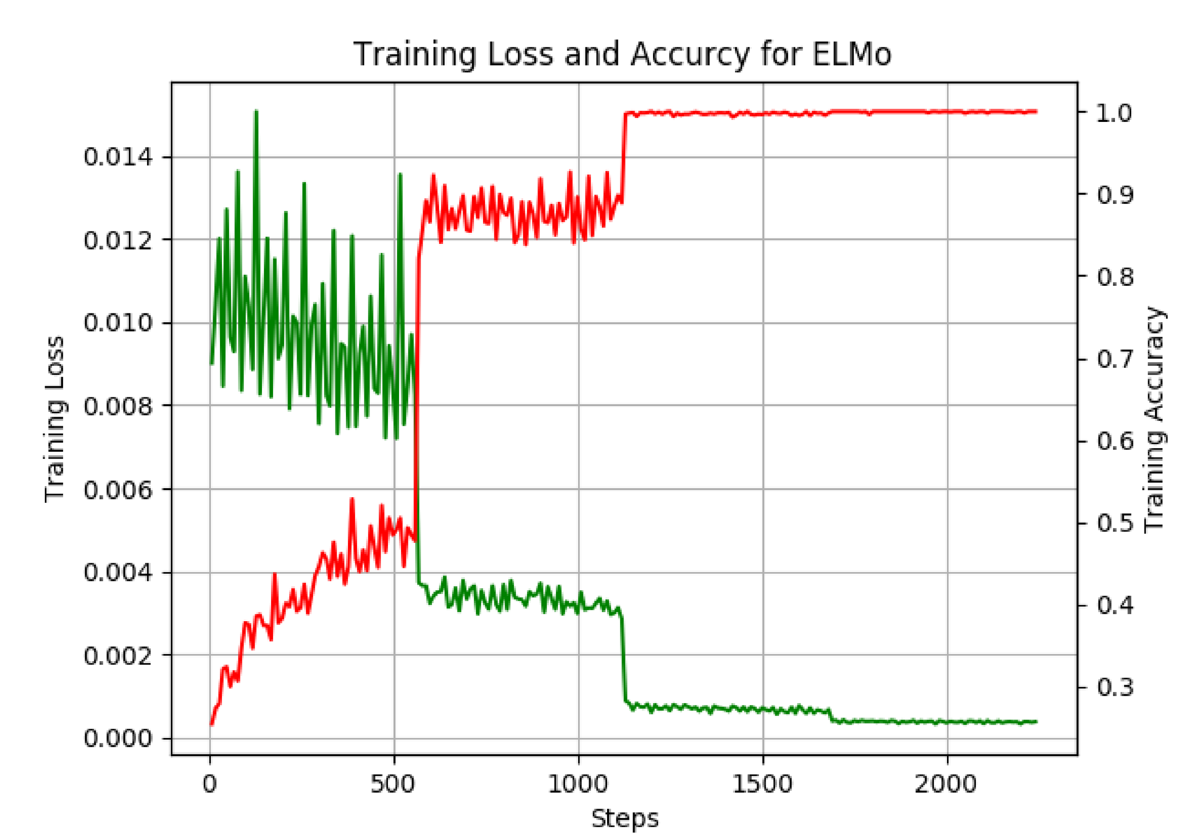

Figure 3. ELMo training loss and accuracy.

Model	CC	CD	DT	JJ	JJR	JJS	NN
BERT	0.78	0.5	0.82	0.71	0.4	0.45	0.67
ELMo	0.52	0.24	0.44	0.33	0.06	0.47	0.28
SenCla	0.67	0.36	0.72	0.61	0.5	0.55	0.57
RB	**RBR**	**TO**	**VB**	**VBG**	**VBN**	**VBZ**	**WDT**
0.73	1.0	0.8	0.85	0.68	0.71	1.0	1.0
0.29	0.21	0.27	0.32	0.22	0.28	0.19	0.14
0.62	0.77	0.68	0.60	0.59	0.59	0.73	0.33

Table 3. Performance By Question Category

BERT performs the best in most categories. Notice that Elmo has better performance in JJS category than BERT, and sentence classification beats BERT on both JJS and JJR. Sentence classification perform not as well on CC, CD and WDT because those categories require contextual information. Elmo performed poorly due to a lack of parameter tuning and the fact that its weighted were trained for next word in sequence prediction.

Future Works

Overfitting of ELMo and TCN models can be alleviated by introducing more datasets. These two models should be trained on single-sentence completion before fine-tuning on CLOTH dataset.
Given the superior performance of BERT, difficult cloze questions can be generated by interpreting BERT output. It is worth investigating human performance on BERT-generated articles.

References

[1] Xie, Qizhe, et al. "Large-scale Cloze Test Dataset Created by Teachers." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
[3] Peters, Matthew, et al. "Deep Contextualized Word Representations." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1. 2018.