自督導式學習的神奇能力

李宏毅
Outline

Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages
Multi-lingual BERT

Training a BERT model by many different languages.
Zero-shot Reading Comprehension

Training on the sentences of 104 languages

Train on English QA training examples

Test on Chinese QA test

Multi-BERT
Zero-shot Reading Comprehension

- English: SQuAD, Chinese: DRCD

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-train</th>
<th>Fine-tune</th>
<th>Test</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>QANet</td>
<td>none</td>
<td>Chinese</td>
<td></td>
<td>66.1</td>
<td>78.1</td>
</tr>
<tr>
<td>BERT</td>
<td>Chinese 104</td>
<td>Chinese</td>
<td>Chinese</td>
<td>82.0</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>languages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Chinese</td>
<td></td>
<td>81.2</td>
<td>88.7</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td></td>
<td></td>
<td>63.3</td>
<td>78.8</td>
</tr>
<tr>
<td></td>
<td>Chinese + English</td>
<td></td>
<td></td>
<td>82.6</td>
<td>90.1</td>
</tr>
</tbody>
</table>

F1 score of Human performance is 93.30%

This work is done by 劉記良、許宗嫄
So many evidences ......

Table 1: NER F1 results on the CoNLL data.

<table>
<thead>
<tr>
<th>Fine-tuning \ Eval</th>
<th>EN</th>
<th>DE</th>
<th>NL</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>90.70</td>
<td>69.74</td>
<td>77.36</td>
<td>73.59</td>
</tr>
<tr>
<td>DE</td>
<td>73.83</td>
<td>82.00</td>
<td>76.25</td>
<td>70.03</td>
</tr>
<tr>
<td>NL</td>
<td>65.46</td>
<td>65.68</td>
<td>89.86</td>
<td>72.10</td>
</tr>
<tr>
<td>ES</td>
<td>65.38</td>
<td>59.40</td>
<td>64.39</td>
<td>87.18</td>
</tr>
</tbody>
</table>

Table 2: POS accuracy on a subset of UD languages.

<table>
<thead>
<tr>
<th>Fine-tuning \ Eval</th>
<th>EN</th>
<th>DE</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>96.82</td>
<td>89.40</td>
<td>85.91</td>
<td>91.60</td>
</tr>
<tr>
<td>DE</td>
<td>83.99</td>
<td>93.99</td>
<td>86.32</td>
<td>88.39</td>
</tr>
<tr>
<td>ES</td>
<td>81.64</td>
<td>88.87</td>
<td>96.71</td>
<td>93.71</td>
</tr>
<tr>
<td>IT</td>
<td>86.79</td>
<td>87.82</td>
<td>91.28</td>
<td>98.11</td>
</tr>
</tbody>
</table>

https://aclanthology.org/P19-1493/

https://aclanthology.org/D19-1077/
Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark

https://sites.research.google/xtreme

40 languages for 9 tasks
Train on English, and test on the rest
Cross-lingual Alignment?

Multi-BERT

深  high  度  est  學  moun  習  tain

跳  jump

兔  rabbit

游  swim

魚  fish
How alignment happens?

- Typical answer

Different languages share some common tokens.

How do you explain Chinese v.s. English?

Code Switching

... DNA 的構造很像螺旋梯 ...
(digits, punctuations)

Intermediate Language?

Language X shares tokens with Chinese and English.
How alignment happens?

https://openreview.net/forum?id=HJeT3yrtDr

<table>
<thead>
<tr>
<th>B-BERT</th>
<th>Train</th>
<th>Test</th>
<th>XNLI</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-es</td>
<td>en</td>
<td>es</td>
<td>72.3</td>
<td>61.9 (±0.8)</td>
</tr>
<tr>
<td>en-enfake-es</td>
<td>enfake</td>
<td></td>
<td>70.9</td>
<td>62.6 (±1.6)</td>
</tr>
<tr>
<td>en-hi</td>
<td>en</td>
<td>hi</td>
<td>60.1</td>
<td>61.6 (±0.7)</td>
</tr>
<tr>
<td>en-enfake-hi</td>
<td>enfake</td>
<td></td>
<td>59.6</td>
<td>62.9 (±0.7)</td>
</tr>
<tr>
<td>en-ru</td>
<td>en</td>
<td>ru</td>
<td>66.4</td>
<td>57.1* (±0.9)</td>
</tr>
<tr>
<td>en-enfake-ru</td>
<td>enfake</td>
<td></td>
<td>65.7</td>
<td>54.2 (±0.7)</td>
</tr>
<tr>
<td>en-enfake</td>
<td>enfake</td>
<td>enfake</td>
<td>78.0</td>
<td>78.9* (±0.7)</td>
</tr>
<tr>
<td>en-enfake</td>
<td>enfake</td>
<td>en</td>
<td>77.5</td>
<td>76.6(±0.8)</td>
</tr>
</tbody>
</table>

English: the cat is a good cat
Fake-English: 甲 乙 天地人 乙
Weird???

If the embedding is language independent ...

How to correctly reconstruct?

There must be language information.

Where is Language?

That is done by 劉記良、許宗嫄、莊永松

Multi-BERT

Reconstruction

Average of Chinese

Average of English

那 有 一 貓

跳 jump

兔 rabbit

游 swim

魚 fish

there is a cat

fish

rabbit

jump

swim
If this is true ...

Average of Chinese

Average of English

Unsupervised token-level translation 😊
Table 1: Unsupervised Token Translation quantitative results using the 10-th layer of BERT.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1 (α=1)</td>
<td>7.53</td>
<td>8.53</td>
<td>5.56</td>
<td>7.96</td>
<td>15.25</td>
<td>7.88</td>
<td>7.34</td>
<td>9.08</td>
<td>5.52</td>
<td>6.34</td>
<td>4.37</td>
<td>6.54</td>
</tr>
<tr>
<td>BLEU-1 (α=2)</td>
<td>8.03</td>
<td>10.24</td>
<td>6.31</td>
<td>7.23</td>
<td>21.51</td>
<td>14.91</td>
<td>7.48</td>
<td>8.52</td>
<td>6.23</td>
<td>7.48</td>
<td>5.38</td>
<td>6.65</td>
</tr>
<tr>
<td>BLEU-1 (α=3)</td>
<td>12.35</td>
<td>10.65</td>
<td>5.35</td>
<td>7.16</td>
<td>15.95</td>
<td>19.13</td>
<td>6.29</td>
<td>12.27</td>
<td>5.74</td>
<td>6.45</td>
<td>6.17</td>
<td>4.73</td>
</tr>
<tr>
<td>convert rate (α=1)</td>
<td>40.2</td>
<td>41.7</td>
<td>61.1</td>
<td>15.3</td>
<td>47.8</td>
<td>62.1</td>
<td>45.2</td>
<td>49.6</td>
<td>29.9</td>
<td>14.7</td>
<td>23.9</td>
<td>30.2</td>
</tr>
<tr>
<td>convert rate (α=2)</td>
<td>74.8</td>
<td>75.7</td>
<td>99.4</td>
<td>97.4</td>
<td>90.0</td>
<td>99.1</td>
<td>67.3</td>
<td>601</td>
<td>83.0</td>
<td>65.6</td>
<td>60.8</td>
<td>97.9</td>
</tr>
<tr>
<td>convert rate (α=3)</td>
<td>95.2</td>
<td>96.3</td>
<td>99.8</td>
<td>100</td>
<td>99.5</td>
<td>100</td>
<td>79.5</td>
<td>73.1</td>
<td>96.6</td>
<td>93.6</td>
<td>91.4</td>
<td>99.7</td>
</tr>
</tbody>
</table>

Input (en) | The girl that can help me is all the way across town. There is no one who can help me.

Ground Truth (zh) | 能帮助我的女孩在小镇的另一边，没有人能帮助我。
| en-zh, α = 1 | 孩，can 来我是all the way across 市。。There 是无人 can help 我。
| en-zh, α = 2 | 孩的的家我是这个人的市。。他是他 人 的到我。
| en-zh, α = 3 | ，的的的他是的的的的，，，：他是他人，的。他。

Unsupervised token-level translation 😊
On the Language Neutrality of Pre-trained Multilingual Representations
It’s not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT
Zero-shot Cross-lingual Transfer

Train on English

Downstream

Multi-BERT

Test on Chinese

Downstream

Multi-BERT

different?
Downstream
Train on English
Multi-BERT

Average of English

Average of Chinese

Train on English
Downstream
Multi-BERT

Test on Chinese
Downstream
Multi-BERT

Similar?
## Experimental Results

### Table 4: POS tagging results

<table>
<thead>
<tr>
<th>Method</th>
<th>ar</th>
<th>bg</th>
<th>de</th>
<th>el</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ru</th>
<th>th</th>
<th>tr</th>
<th>ur</th>
<th>vi</th>
<th>zh</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>53.8</td>
<td>85.4</td>
<td>86.2</td>
<td>81.1</td>
<td>86.1</td>
<td>42.9</td>
<td>66.8</td>
<td>85.5</td>
<td>41.7</td>
<td>68.6</td>
<td>56.3</td>
<td>53.8</td>
<td>61.8</td>
<td>66.9</td>
</tr>
<tr>
<td>Zero-mean</td>
<td>54.3</td>
<td>86.1</td>
<td><strong>86.6</strong></td>
<td><strong>81.8</strong></td>
<td>86.6</td>
<td>43.7</td>
<td>68.1</td>
<td><strong>86.5</strong></td>
<td>41.6</td>
<td><strong>69.7</strong></td>
<td>56.6</td>
<td>53.4</td>
<td>62.5</td>
<td><strong>67.5</strong></td>
</tr>
<tr>
<td>MDS</td>
<td>54.2</td>
<td><strong>86.4</strong></td>
<td>86.5</td>
<td>81.5</td>
<td><strong>86.8</strong></td>
<td><strong>43.9</strong></td>
<td><strong>68.9</strong></td>
<td>86.4</td>
<td><strong>44.2</strong></td>
<td>69.4</td>
<td><strong>57.1</strong></td>
<td>52.4</td>
<td><strong>63.0</strong></td>
<td><strong>67.8</strong></td>
</tr>
</tbody>
</table>

### Table 5: Dependency parsing results. Numbers are Labeled Attachment Score (LAS).

<table>
<thead>
<tr>
<th>Method</th>
<th>ar</th>
<th>bg</th>
<th>de</th>
<th>el</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ru</th>
<th>th</th>
<th>tr</th>
<th>ur</th>
<th>vi</th>
<th>zh</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>28.2</td>
<td>70.7</td>
<td><strong>74.0</strong></td>
<td><strong>71.6</strong></td>
<td>72.1</td>
<td>74.8</td>
<td>35.3</td>
<td>69.0</td>
<td>30.8</td>
<td>32.9</td>
<td>28.3</td>
<td><strong>37.8</strong></td>
<td><strong>35.4</strong></td>
<td>50.8</td>
</tr>
<tr>
<td>Zero-mean</td>
<td>28.2</td>
<td><strong>71.0</strong></td>
<td>73.4</td>
<td>71.4</td>
<td>72.2</td>
<td><strong>75.7</strong></td>
<td>36.3</td>
<td><strong>69.3</strong></td>
<td>32.5</td>
<td><strong>34.6</strong></td>
<td>28.6</td>
<td>37.0</td>
<td>35.2</td>
<td><strong>51.2</strong></td>
</tr>
<tr>
<td>MDS</td>
<td>28.0</td>
<td>70.8</td>
<td>73.7</td>
<td>71.1</td>
<td><strong>72.2</strong></td>
<td>75.3</td>
<td><strong>36.5</strong></td>
<td>68.8</td>
<td>30.4</td>
<td>34.2</td>
<td><strong>29.0</strong></td>
<td>35.6</td>
<td>35.0</td>
<td>50.8</td>
</tr>
</tbody>
</table>
You just have to ask ......

It’s not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT

<table>
<thead>
<tr>
<th></th>
<th>@1</th>
<th>@10</th>
<th>@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.036</td>
<td>0.244</td>
<td>0.575</td>
</tr>
<tr>
<td>Analogies</td>
<td>0.105</td>
<td>0.463</td>
<td>0.737</td>
</tr>
<tr>
<td>Template</td>
<td>0.449</td>
<td>0.703</td>
<td>0.845</td>
</tr>
</tbody>
</table>

Multi-BERT

The word apple in French is [MASE]
Outline

Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages
Why does BERT work?

• Applying BERT to **protein, DNA, music classification**

[DNA sequence image]

class DNA sequence

---


This work is done by 高瑋聰
Why does BERT work?

This work is done by 高瑋聰


DNA sequence

<table>
<thead>
<tr>
<th>A</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>you</td>
</tr>
<tr>
<td>C</td>
<td>he</td>
</tr>
<tr>
<td>G</td>
<td>she</td>
</tr>
</tbody>
</table>

BERT

[CLS]

we → she → we → he

Random initialization

pre-train on English

Init by pre-train

Linear

class
Why does BERT work?

- Applying BERT to **protein, DNA, music classification**

<table>
<thead>
<tr>
<th></th>
<th>Protein</th>
<th></th>
<th></th>
<th>DNA</th>
<th></th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>localization</td>
<td>stability</td>
<td>fluorescence</td>
<td>H3</td>
<td>H4</td>
<td>H3K9ac</td>
</tr>
<tr>
<td>specific</td>
<td>69.0</td>
<td>76.0</td>
<td>63.0</td>
<td>87.3</td>
<td>87.3</td>
<td>79.1</td>
</tr>
<tr>
<td>BERT</td>
<td>64.8</td>
<td>74.5</td>
<td>63.7</td>
<td>83.0</td>
<td>86.2</td>
<td>78.3</td>
</tr>
<tr>
<td>re-emb</td>
<td>63.3</td>
<td>75.4</td>
<td>37.3</td>
<td>78.5</td>
<td>83.7</td>
<td>76.3</td>
</tr>
<tr>
<td>rand</td>
<td>58.6</td>
<td>65.8</td>
<td>27.5</td>
<td>75.6</td>
<td>66.5</td>
<td>72.8</td>
</tr>
</tbody>
</table>

This work is done by 高瑋聰
(a) Synthetic GLUE (8 tasks)  
(b) Protein (3 tasks)  
(c) DNA (4 tasks)  
(d) music (1 task)  

<table>
<thead>
<tr>
<th>Method</th>
<th>Flu.</th>
<th>Stab.</th>
<th>Loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT - PLUS</td>
<td>0.729</td>
<td>0.634</td>
<td>0.504</td>
</tr>
<tr>
<td>BERT - random</td>
<td>0.598</td>
<td>0.545</td>
<td>0.362</td>
</tr>
<tr>
<td>PLUS - random</td>
<td>0.461</td>
<td>0.405</td>
<td>0.322</td>
</tr>
<tr>
<td>random - random</td>
<td>0.434</td>
<td>0.388</td>
<td>0.387</td>
</tr>
</tbody>
</table>
**Optimization**

(b) fluorescence

**Generalization**

(a) H3K9ac

(b) localization
Self-supervised Model as Universal Computation Engine

Performance on Multimodal Sequence Benchmarks
這些發現有甚麼用？

Speech Question Answering
Audio Story:  🎧 (The original story is 5 min long.)

Question: “What is a possible origin of Venus’ clouds?”

Choices:

(A) gases released as a result of volcanic activity
(B) chemical reactions caused by high surface temperatures
(C) bursts of radio energy from the plane's surface
(D) strong winds that blow dust into the atmosphere
SQuAD-style Spoken QA

• Link: https://github.com/chiahsuan156/ODSQA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>QA-pairs</th>
<th>Hours</th>
<th>M-spkrs</th>
<th>F-spkrs</th>
<th>WER-D(%)</th>
<th>WER-Q(%)</th>
<th>Avg D Len</th>
<th>AvgQ Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODSQA</td>
<td>3654</td>
<td>25.28</td>
<td>7</td>
<td>13</td>
<td>19.11</td>
<td>18.57</td>
<td>428</td>
<td>22</td>
</tr>
<tr>
<td>DRCD-TTS</td>
<td>16746</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>33.63</td>
<td>--</td>
<td>332</td>
<td>20</td>
</tr>
</tbody>
</table>

SPOKEN OPEN-DOMAIN QUESTION ANSWERING DATASET

https://arxiv.org/abs/1804.00320
SQuAD-style Spoken QA

- Link: [https://github.com/chiahsuan156/ODSQA](https://github.com/chiahsuan156/ODSQA)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>QA-pairs</th>
<th>Hours</th>
<th>M-spkrs</th>
<th>F-spkrs</th>
<th>WER-D(%)</th>
<th>WER-Q(%)</th>
<th>Avg D Len</th>
<th>AvgQ Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODSQA</td>
<td>3654</td>
<td>25.28</td>
<td>7</td>
<td>13</td>
<td>19.11</td>
<td>18.57</td>
<td>428</td>
<td>22</td>
</tr>
<tr>
<td>DRCD-TTS</td>
<td>16746</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>33.63</td>
<td>--</td>
<td>332</td>
<td>20</td>
</tr>
</tbody>
</table>

**SPOKEN OPEN-DOMAIN QUESTION ANSWERING DATASET**

**SOD QA**

**OPEN-DOMAIN SPOKEN QUESTION ANSWERING DATASET**

**ODS QA**

[Lee, et al., SLT'18]
Towards End-to-end

Speech Question Answering

Training data

$q_1$ $d_1$ ans span

$q_2$ $d_2$ ans span

$q_3$ $d_3$ ans span

(no speech transcription)

Can we train an end-to-end speech QA model?
Speech Question Answering

It does not work ..... No semantic information

Pre-trained Speech Model

[CLS] question [SEP] document

Find Ans Span

start end
It does not work ....

Pre-trained Speech Model

Find Ans Span

start

end

Pre-trained Speech Model

[CLS]

question

[SEP]

document
No pre-train: 6.12 F1 score → Pre-training on text: 54.22 F1 score

Pre-trained Speech Model

BERT (pre-trained on text)

Pre-trained Speech Model

[CLS] [SEP] question document

VQ

Find Ans Span

start end
Speech Question Answering

<table>
<thead>
<tr>
<th>Embedding Assignment</th>
<th>FF1</th>
<th>AOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most frequent</td>
<td>54.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Least frequent</td>
<td>46.9</td>
<td>41.7</td>
</tr>
<tr>
<td>Random</td>
<td>51.7</td>
<td>46.2</td>
</tr>
<tr>
<td>Re-init</td>
<td>8.9</td>
<td>7.2</td>
</tr>
<tr>
<td>Scratch (baseline)</td>
<td>6.1</td>
<td>4.9</td>
</tr>
</tbody>
</table>

https://arxiv.org/abs/2203.04911
This work is done by 林冠廷
Outline

- Case Study: BERT
- Story 1: Cross-lingual
- Story 2: Cross-discipline
- Story 3: Pre-training without Human Languages
Pre-training on Artificial Data

By defining different rules, we know what are the key factors for the success of pre-training.

Token generation by rules

Pre-train

Fine-tune

English downstream tasks

BERT

1 2 1 1 7 7 ......
3 9 3 6 5 4 ......
Stage 1
L1 MLM pre-train

Stage 2
GLUE fine-tune

Stage 3
GLUE Testing

Token ID | Input Token
--- | ---
0 | is
1 | it
2 | very
3 | good
4 | movie
5 | a
6 | bad
7 | ,
Pre-training on Artificial Data

GLUE score improvement vs. train from scratch

This work is done by 姜成翰

https://arxiv.org/abs/2109.03537
Pre-training on Artificial Data

GLUE score improvement vs. train from scratch

- Pre-training on random generated words yields the same performance as training from scratch.
- Data plays the role.

https://arxiv.org/abs/2109.03537
This work is done by 姜成翰
Data is critical ......

Learn more from robust wav2vec 2.0
https://arxiv.org/abs/2104.01027

<table>
<thead>
<tr>
<th></th>
<th>continual</th>
<th>clean</th>
<th>IC (Acc)</th>
<th>ER (Acc)</th>
<th>KS (Acc)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>m+g+r</td>
<td>fsd50k</td>
<td>m+g+r</td>
</tr>
<tr>
<td>(a) baseline</td>
<td>-</td>
<td>99.47</td>
<td>96.94</td>
<td>97.47</td>
<td>63.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>continual</th>
<th>clean</th>
<th>SID (Acc)</th>
<th>ASR (WER)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>m+g+r</td>
<td>fsd50k</td>
</tr>
<tr>
<td>(a) baseline</td>
<td>-</td>
<td>84.97</td>
<td>65.51</td>
<td>77.61</td>
</tr>
</tbody>
</table>
Data is critical ......

Continuously train with noisy data

Pre-trained Model
HuBERT (Librispeech)

Labeled Data
e.g., clean

Testing Data
e.g., noisy

Learn more from robust wav2vec 2.0
https://arxiv.org/abs/2104.01027

<table>
<thead>
<tr>
<th></th>
<th>continual</th>
<th>clean</th>
<th>IC (Acc) m+g+r fsd50k</th>
<th>ER (Acc) m+g+r fsd50k</th>
<th>KS (Acc) m+g+r fsd50k</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) baseline</td>
<td>-</td>
<td>99.47</td>
<td>96.94</td>
<td>97.47</td>
<td>63.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) w/o DAT</td>
<td>libri 100hr mgr</td>
<td>99.45</td>
<td>98.63</td>
<td>97.94</td>
<td>64.42</td>
</tr>
<tr>
<td>(c) w/o DAT</td>
<td>libri 960hr mgr</td>
<td>99.39</td>
<td>98.84</td>
<td>97.89</td>
<td>67.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>clean</th>
<th>SID (Acc) m+g+r fsd50k</th>
<th>ASR (WER)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>continual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) baseline</td>
<td>-</td>
<td>84.97</td>
<td>65.51</td>
<td></td>
</tr>
<tr>
<td>(b) w/o DAT</td>
<td>libri 100hr mgr</td>
<td>87.02</td>
<td>70.91</td>
<td></td>
</tr>
<tr>
<td>(c) w/o DAT</td>
<td>libri 960hr mgr</td>
<td>86.40</td>
<td>74.46</td>
<td></td>
</tr>
</tbody>
</table>
Pre-training on Artificial Data

GLUE score improvement vs. train from scratch

- Pre-training on random generated words yields the same performance as training from scratch.
- Data plays the role.
- Pre-training from the data generated by unigram or bigram LMs helps a little.
Pre-training on Artificial Data

- The sentences generated by very simple rules can lead to good pre-trained models.
- All the words in the generated sentences are paired.

GLUE score improvement vs. train from scratch

Also refer to:
Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models

Pre-training on Artificial Data

- The sentences generated by very simple rules can lead to good pre-trained models.
- All the words in the generated sentences are paired.
- Shuffle

GLUE score improvement vs. train from scratch

This work is done by 姜成翰

https://arxiv.org/abs/2109.03537
Concluding Remarks

Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages