

# Using the Transformer

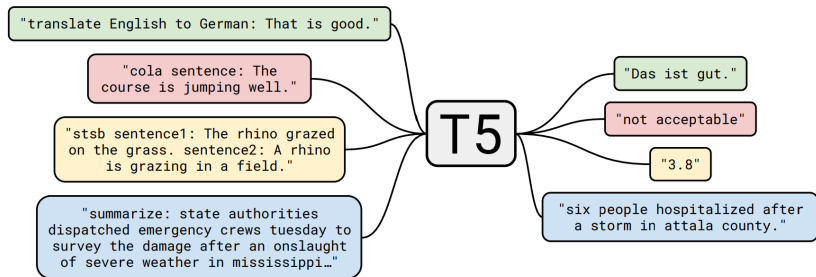
## Text-to-Text Transfer Transformer



### Learning goals

- shortcomings of BERT & Co.
- everything as text-to-text
- Dynamic Masking

# REVISITING TEXT-TO-TEXT TASKS



► Source: Raffel et al., 2020

## Ingredients:

- `<task prefix>` for each task
- Concatenation with the input
- Output label / score formatted as string

# TEXT-TO-TEXT TRANSFER TRANSFORMER

## Short summary:

- In short: T5 (bc of the five Ts)
- A complete encoder-decoder Transformer architecture
- **Question:** Why is this important?
- Relative positional embeddings
- All tasks reformulated as text-to-text tasks
  - Probably the most important innovation of this work
- From BERT-size (110M parameters) up to 11 Billion parameters
- Paradigm shift from *Sequential Transfer Learning* towards true *Multi-Task Learning*

► Nice Animation (similar to figure before)

# T5 – INPUT AND OUTPUT FORMAT

## Input Side:

- SentencePiece framework w/ WordPiece tokens
  - Add task-specific (text) prefix to the original input
  - Choice of the prefix: Hyperparameter!!
- Changing this had limited impact
- No extensive experiments performed by the authors

## Output Side:

- Output as a word or a piece of text (also numerical similarity scores)
- If output not present in set of potential alternatives, prediction considered as wrong (many other possible ways; this is not trivial)

# ADD-ON: DISTINCTION TO PROMPTING

## Adding task-specific (text) prefix:

- Add task-specific (text) prefix to the original input
  - Model is fine-tuned on samples prepended with this prefix
- It learns to recognize what it is expected to do when encountering a specific prefix at test time
- Probably because of this: limited impact found by the authors

## Prompting:

- Refers to just querying a trained w/o fine-tuning it (cf. next chapter)
- Paradigm of Few-/Zero-Shot Learning
- This is found to have a *huge* impact on model performance

# PRE-TRAINING T5

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

Baseline objective (Source: Raffel et al., 2019)

- 1 Mark spans in input sequence for corruptions
- 2 Replace tokens with sentinel tokens
- 3 Predict sentinel tokens and replaced tokens

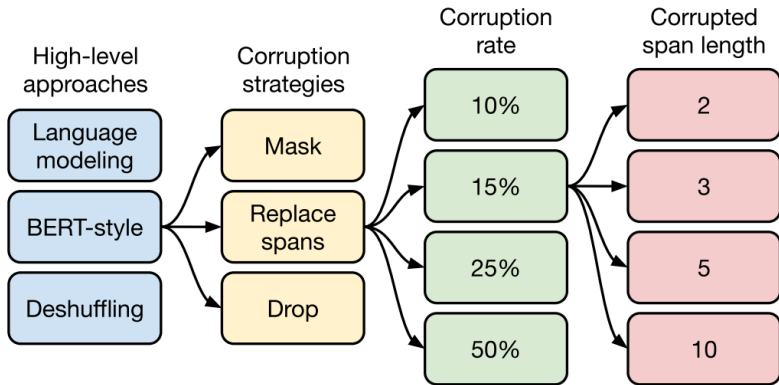
# PRE-TRAINING OBJECTIVES

- Authors experimented with different objectives
- Most of them rely in some way on MLM

| Objective                                       | Inputs  | Targets                                     |
|---|---|---|
| Prefix language modeling                        | Thank you for inviting                                  | me to your party last week .                |
| BERT-style <a href="#">Devlin et al. (2018)</a> | Thank you <M> <M> me to your party apple week .         | <i>(original text)</i>                      |
| Deshuffling                                     | party me for your to . last fun you inviting week Thank | <i>(original text)</i>                      |
| MASS-style <a href="#">Song et al. (2019)</a>   | Thank you <M> <M> me to your party <M> week .           | <i>(original text)</i>                      |
| I.i.d. noise, replace spans                     | Thank you <X> me to your party <Y> week .               | <X> for inviting <Y> last <Z>               |
| I.i.d. noise, drop tokens                       | Thank you me to your party week .                       | for inviting last                           |
| Random spans                                    | Thank you <X> to <Y> week .                             | <X> for inviting me <Y> your party last <Z> |

Source: Raffel et al. (2019)

# PRE-TRAINING OBJECTIVES



Source: Raffel et al. (2019)



# THE COLOSSAL CLEAN CRAWLED CORPUS (C4)

- Effort to measure the effect of quality, characteristics & size of the pre-training resources
- Common Crawl as basis, careful cleaning and filtering for English language
- Orders of magnitude larger (750GB) compared to commonly used corpora

# THE COLOSSAL CLEAN CRAWLED CORPUS (C4)

## Experiments (with respect to C4)

| Data set        | Size  | GLUE         | CNNDM        | SQuAD        | SGLUE        | EnDe         | EnFr         | EnRo         |
|-----------------|-------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| ★ C4            | 745GB | 83.28        | <b>19.24</b> | 80.88        | 71.36        | <b>26.98</b> | <b>39.82</b> | <b>27.65</b> |
| C4, unfiltered  | 6.1TB | 81.46        | 19.14        | 78.78        | 68.04        | 26.55        | 39.34        | 27.21        |
| RealNews-like   | 35GB  | <b>83.83</b> | <b>19.23</b> | 80.39        | 72.38        | <b>26.75</b> | <b>39.90</b> | <b>27.48</b> |
| WebText-like    | 17GB  | <b>84.03</b> | <b>19.31</b> | <b>81.42</b> | 71.40        | <b>26.80</b> | <b>39.74</b> | <b>27.59</b> |
| Wikipedia       | 16GB  | 81.85        | <b>19.31</b> | 81.29        | 68.01        | <b>26.94</b> | 39.69        | <b>27.67</b> |
| Wikipedia + TBC | 20GB  | 83.65        | <b>19.28</b> | <b>82.08</b> | <b>73.24</b> | <b>26.77</b> | 39.63        | <b>27.57</b> |

| Number of tokens | Repeats | GLUE         | CNNDM        | SQuAD        | SGLUE        | EnDe         | EnFr         | EnRo         |
|------------------|---------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| ★ Full data set  | 0       | <b>83.28</b> | <b>19.24</b> | <b>80.88</b> | <b>71.36</b> | <b>26.98</b> | <b>39.82</b> | <b>27.65</b> |
| 2 <sup>29</sup>  | 64      | <b>82.87</b> | <b>19.19</b> | <b>80.97</b> | <b>72.03</b> | <b>26.83</b> | <b>39.74</b> | <b>27.63</b> |
| 2 <sup>27</sup>  | 256     | 82.62        | <b>19.20</b> | 79.78        | 69.97        | <b>27.02</b> | <b>39.71</b> | 27.33        |
| 2 <sup>25</sup>  | 1,024   | 79.55        | 18.57        | 76.27        | 64.76        | 26.38        | 39.56        | 26.80        |
| 2 <sup>23</sup>  | 4,096   | 76.34        | 18.33        | 70.92        | 59.29        | 26.37        | 38.84        | 25.81        |

Source: Raffel et al. (2019)

# T5 - EXHAUSTIVE EXPERIMENTS

## Performed experiments with respect to ..

- .. *architecture, size & objective*
  - enc, dec, enc-dec
  - Between 60M and 11B parameters
- .. *details of the Denoising objective (which was found to work best)*
- .. *fine-tuning methods*
  - Adapter layers
  - Gradual Unfreezing (cf. ULMFiT)
- .. *Multi-task learning strategies*
  - Examples-proportional mixing
  - Temperature-scaled mixing

# BENCHMARK RESULTS

| Model         | GLUE<br>Average   | CoLA<br>Matthew's | SST-2<br>Accuracy | MRPC<br>F1              | MRPC<br>Accuracy        | STS-B<br>Pearson  | STS-B<br>Spearman |
|---------------|-------------------|-------------------|-------------------|-------------------------|-------------------------|-------------------|-------------------|
| Previous best | 89.4 <sup>a</sup> | 69.2 <sup>b</sup> | 97.1 <sup>a</sup> | <b>93.6<sup>b</sup></b> | <b>91.5<sup>b</sup></b> | 92.7 <sup>b</sup> | 92.3 <sup>b</sup> |
| T5-Small      | 77.4              | 41.0              | 91.8              | 89.7                    | 86.6                    | 85.6              | 85.0              |
| T5-Base       | 82.7              | 51.1              | 95.2              | 90.7                    | 87.5                    | 89.4              | 88.6              |
| T5-Large      | 86.4              | 61.2              | 96.3              | 92.4                    | 89.9                    | 89.9              | 89.2              |
| T5-3B         | 88.5              | 67.1              | 97.4              | 92.5                    | 90.0                    | 90.6              | 89.8              |
| T5-11B        | <b>90.3</b>       | <b>71.6</b>       | <b>97.5</b>       | 92.8                    | 90.4                    | <b>93.1</b>       | <b>92.8</b>       |

| Model         | QQP<br>F1         | QQP<br>Accuracy         | MNLI-m<br>Accuracy | MNLI-mm<br>Accuracy | QNLI<br>Accuracy        | RTE<br>Accuracy   | WNLI<br>Accuracy  |
|---------------|-------------------|-------------------------|--------------------|---------------------|-------------------------|-------------------|-------------------|
| Previous best | 74.8 <sup>c</sup> | <b>90.7<sup>b</sup></b> | 91.3 <sup>a</sup>  | 91.0 <sup>a</sup>   | <b>99.2<sup>a</sup></b> | 89.2 <sup>a</sup> | 91.8 <sup>a</sup> |
| T5-Small      | 70.0              | 88.0                    | 82.4               | 82.3                | 90.3                    | 69.9              | 69.2              |
| T5-Base       | 72.6              | 89.4                    | 87.1               | 86.2                | 93.7                    | 80.1              | 78.8              |
| T5-Large      | 73.9              | 89.9                    | 89.9               | 89.6                | 94.8                    | 87.2              | 85.6              |
| T5-3B         | 74.4              | 89.7                    | 91.4               | 91.2                | 96.3                    | 91.1              | 89.7              |
| T5-11B        | <b>75.1</b>       | 90.6                    | <b>92.2</b>        | <b>91.9</b>         | 96.9                    | <b>92.8</b>       | <b>94.5</b>       |

Results on GLUE (Source: Raffel et al., 2019)

# BENCHMARK RESULTS

| Model         | SQuAD<br>EM       | SQuAD<br>F1       | SuperGLUE<br>Average | BoolQ<br>Accuracy | CB<br>F1          | CB<br>Accuracy    | COPA<br>Accuracy  |
|---------------|-------------------|-------------------|----------------------|-------------------|-------------------|-------------------|-------------------|
| Previous best | 90.1 <sup>a</sup> | 95.5 <sup>a</sup> | 84.6 <sup>d</sup>    | 87.1 <sup>d</sup> | 90.5 <sup>d</sup> | 95.2 <sup>d</sup> | 90.6 <sup>d</sup> |
| T5-Small      | 79.10             | 87.24             | 63.3                 | 76.4              | 56.9              | 81.6              | 46.0              |
| T5-Base       | 85.44             | 92.08             | 76.2                 | 81.4              | 86.2              | 94.0              | 71.2              |
| T5-Large      | 86.66             | 93.79             | 82.3                 | 85.4              | 91.6              | 94.8              | 83.4              |
| T5-3B         | 88.53             | 94.95             | 86.4                 | 89.9              | 90.3              | 94.4              | 92.0              |
| T5-11B        | <b>91.26</b>      | <b>96.22</b>      | <b>88.9</b>          | <b>91.2</b>       | <b>93.9</b>       | <b>96.8</b>       | <b>94.8</b>       |

| Model         | MultiRC<br>F1a    | MultiRC<br>EM     | ReCoRD<br>F1      | ReCoRD<br>Accuracy | RTE<br>Accuracy   | WiC<br>Accuracy   | WSC<br>Accuracy   |
|---------------|-------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|
| Previous best | 84.4 <sup>d</sup> | 52.5 <sup>d</sup> | 90.6 <sup>d</sup> | 90.0 <sup>d</sup>  | 88.2 <sup>d</sup> | 69.9 <sup>d</sup> | 89.0 <sup>d</sup> |
| T5-Small      | 69.3              | 26.3              | 56.3              | 55.4               | 73.3              | 66.9              | 70.5              |
| T5-Base       | 79.7              | 43.1              | 75.0              | 74.2               | 81.5              | 68.3              | 80.8              |
| T5-Large      | 83.3              | 50.7              | 86.8              | 85.9               | 87.8              | 69.3              | 86.3              |
| T5-3B         | 86.8              | 58.3              | 91.2              | 90.4               | 90.7              | 72.1              | 90.4              |
| T5-11B        | <b>88.1</b>       | <b>63.3</b>       | <b>94.1</b>       | <b>93.4</b>        | <b>92.5</b>       | <b>76.9</b>       | <b>93.8</b>       |

Results on SQUAD and S-GLUE (Source: Raffel et al., 2019)

# BENCHMARK RESULTS

| Model         | WMT EnDe<br>BLEU        | WMT EnFr<br>BLEU        | WMT EnRo<br>BLEU        | CNN/DM<br>ROUGE-1  | CNN/DM<br>ROUGE-2  | CNN/DM<br>ROUGE-L  |
|---------------|-------------------------|-------------------------|-------------------------|--------------------|--------------------|--------------------|
| Previous best | <b>33.8<sup>e</sup></b> | <b>43.8<sup>e</sup></b> | <b>38.5<sup>f</sup></b> | 43.47 <sup>g</sup> | 20.30 <sup>g</sup> | 40.63 <sup>g</sup> |
| T5-Small      | 26.7                    | 36.0                    | 26.8                    | 41.12              | 19.56              | 38.35              |
| T5-Base       | 30.9                    | 41.2                    | 28.0                    | 42.05              | 20.34              | 39.40              |
| T5-Large      | 32.0                    | 41.5                    | 28.1                    | 42.50              | 20.68              | 39.75              |
| T5-3B         | 31.8                    | 42.6                    | 28.2                    | 42.72              | 21.02              | 39.94              |
| T5-11B        | 32.1                    | 43.4                    | 28.1                    | <b>43.52</b>       | <b>21.55</b>       | <b>40.69</b>       |

Results on MT/Summarization Benchmarks (Source: Raffel et al., 2019)

# T5 - EXHAUSTIVE EXPERIMENTS

## Conclusions

- Encoder-decoder architecture works best in this "text-to-text" setting
- More data, larger models & ensembling all boost the performance
  - Larger models trained for fewer steps better than smaller models on more data
  - Ensembling: Using same base pre-trained models worse than complete separate model ensembles
- Different denoising objectives work similarly well
- Updating *all* model parameters during fine-tuning works best (but expensive)