SPONSORED BY THE

Federal Ministry of Education and Research

EffEval: A Comprehensive Evaluation of Efficiency for MT Evaluation Metrics

Daniil Larionov[1], Jens Grünwald[2], Christoph Leiter[1] and Steffen Eger[1]
[1] Natural Language Learning Group, Bielefeld University
[2] Department of Computer Science, Technical University of Darmstadt
daniil.larionov@uni-bielefeld.de

Main Contributions

We study 3 different aspects of computational efficiency in application to MT evaluation metrics.

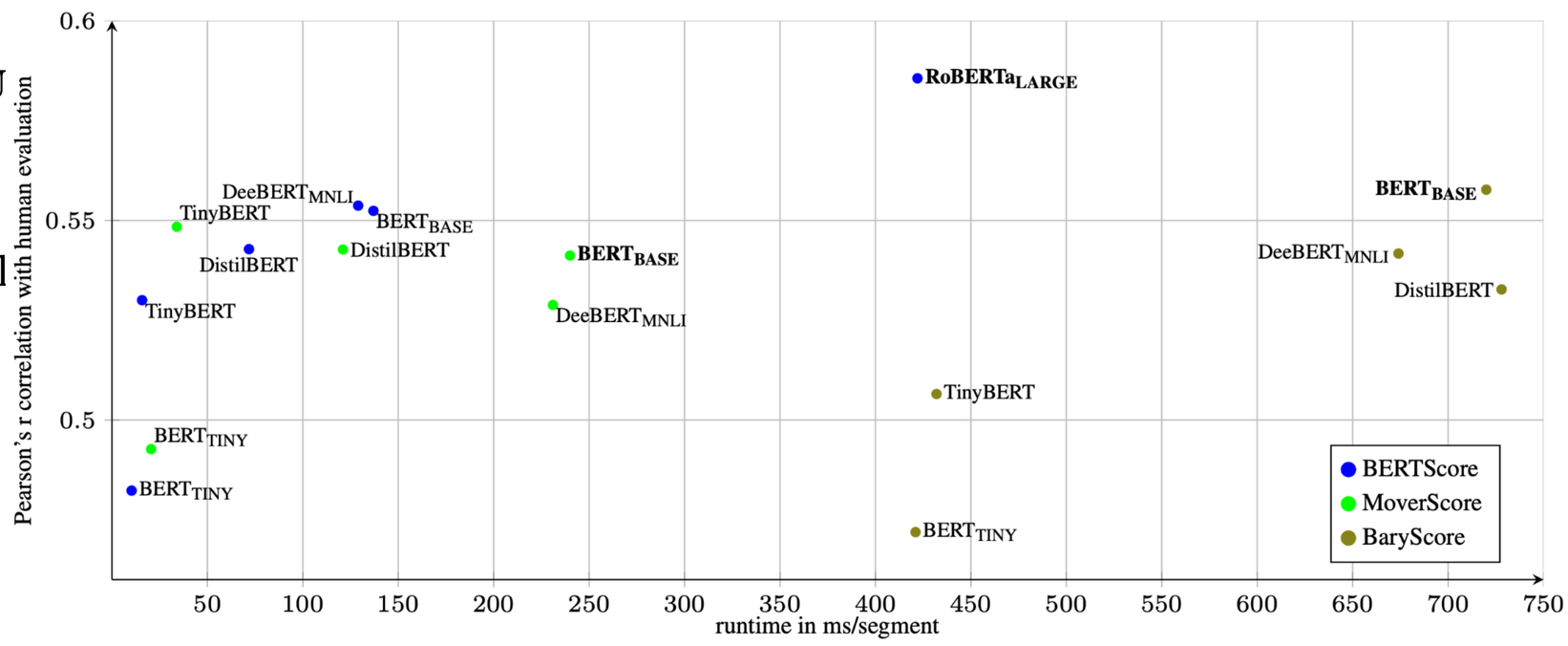
1. We replace computationally-heavy transformer models with their light-weight alternatives for metrics like **BERTScore**, **MoverScore**, **BaryScore**.
2. We switch from costly alignment techniques (Word Mover Distance - WMD) to their approximations in **MoverScore**
3. We train **COMET**-like models efficiently with adapters

Motivation

- Inefficient metrics (like BERTScore with RoBERTa-Large) require expensive hardware to run in reasonable time, which prevents **under-resourced practitioners** from using it.
- Even with good hardware it takes too much time and energy: **71** hours for BERTScore to evaluate (30k segments × 5 language pairs × 50 MT systems)
- Metrics have loads of other applications which would benefit from an efficient option: **RL reward functions**, **mined data filtering**, **online re-ranking**

1. Efficient Transformers

- Evaluation datasets: WMT 15/16/21
- Measured runtime in ms/segment, on GPU and CPU, averaged across 3 runs.
- Evaluated models:
  - RoBERTa-Large - Baseline
  - BERT-Base - Smaller model
  - DistilBERT - Distilled
  - TinyBERT - Adv. Distilled
  - BERT-Tiny miniatures - BERT
  - DeeBERT exiting - Dyn. early
- Best tradeoff between quality/efficiency - **TinyBERT**



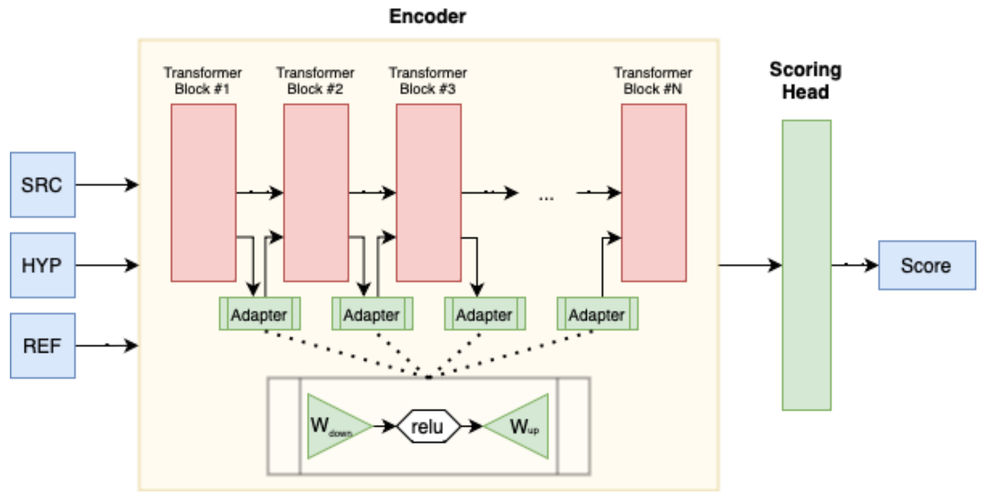
2. Approximations of Word Mover Distance

- MoverScore uses **WMD** as measure of distance between two texts.
- WMD calculates minimum cumulative distance that words from one text needs to travel to match the other text. Complexity is **$O(p^3 \log p)$.**
- Word Centroid Distance (**WCD**) and Relaxed Word Mover Distance (**RWMD**) are two fast approximations of WMD.

Step	WMD	WCD	RWMD
get BERT embeddings	285.499	287.915	291.122
calculate distance matrix	0.829	0.005	0.782
calculate distance	5.602	0.616	0.449

3. COMET + Adapters = ❤️?

- COMET is a trained MT evaluation metric, based on XLM-RoBERTa-Large. It is quite large and takes days to train fully. **An ideal case for adapters!**
- We tested various adapter configurations: Pfeiffer et al, Houlsby et al, Parallel adapter, Compacter, (IA)3. Measured **Mem.** in MB/Token, **Fwd.** and **Bwd.** speed in Tokens/Second and Kendall-tau correlation on WMT21
- Along with large COMET we also tested smaller COMETINHO.
- Results:
  - Adapters **decrease** GPU RAM usage for training and **improve** backward pass speed
  - Models with adapter can **outperform** the ones without them!
  - Simpler adapters works better



Config	Mem.↓	Fwd.↑	Bwd.↑	τ ↑
pfeiffer	4.88	5123	**4808**	0.273
parallel	4.97	5128	4525	**0.289**
houlsby	4.87	4607	4036	0.273
compacter	**4.80**	3649	3049	0.269
$(IA)^3$	5.76	**5195**	4712	0.268
no adapters	7.32	6247	2238	0.275
reference	-	-	-	**0.290**

Config	Mem.↓	Fwd.↑	Bwd.↑	τ ↑
pfeiffer	0.770	25499	25774	0.252
parallel	**0.741**	26109	**26113**	**0.252**
houlsby	0.769	23746	21678	0.252
compacter	0.776	18382	15671	0.243
$(IA)^3$	0.997	**27075**	24804	0.248
no adapters	1.012	31836	18941	0.243
reference	-	-	-	**0.241**

Github