BERT-Based Evaluation of Text Generation Systems

Steffen Eger

eger@aiphes.tu-darmstadt.de

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

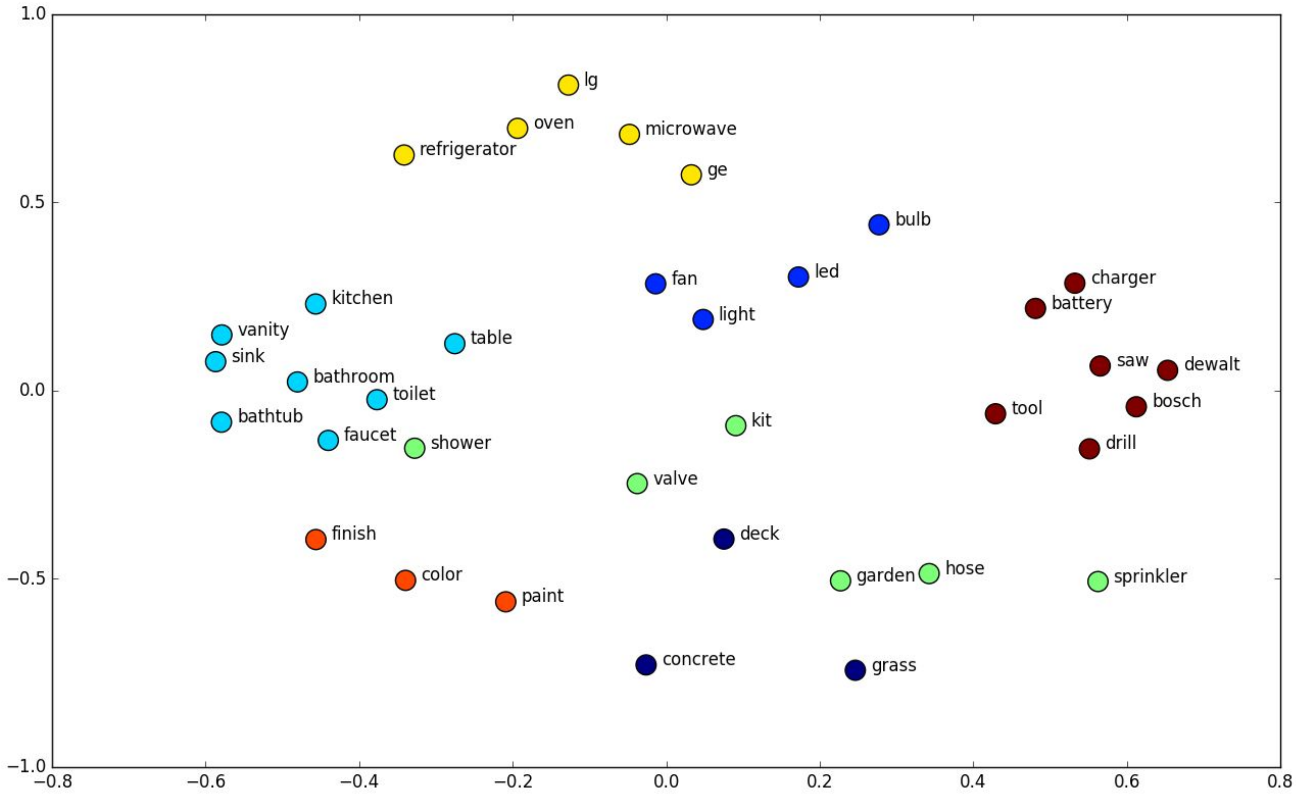
This talk has three parts

- Contextualized Embeddings (very briefly)
- Referenced-based Evaluation with BERT
- Reference-free Evaluation with BERT

(Static) Embeddings

- Vector representations derived from neural networks were popularized by Word2Vec (2013-2014)
- Many extensions: Dependency Based Embeddings (2015), FastText (2017), Multilingual Embeddings (2014-2018)
- All of them had one drawback: they were static

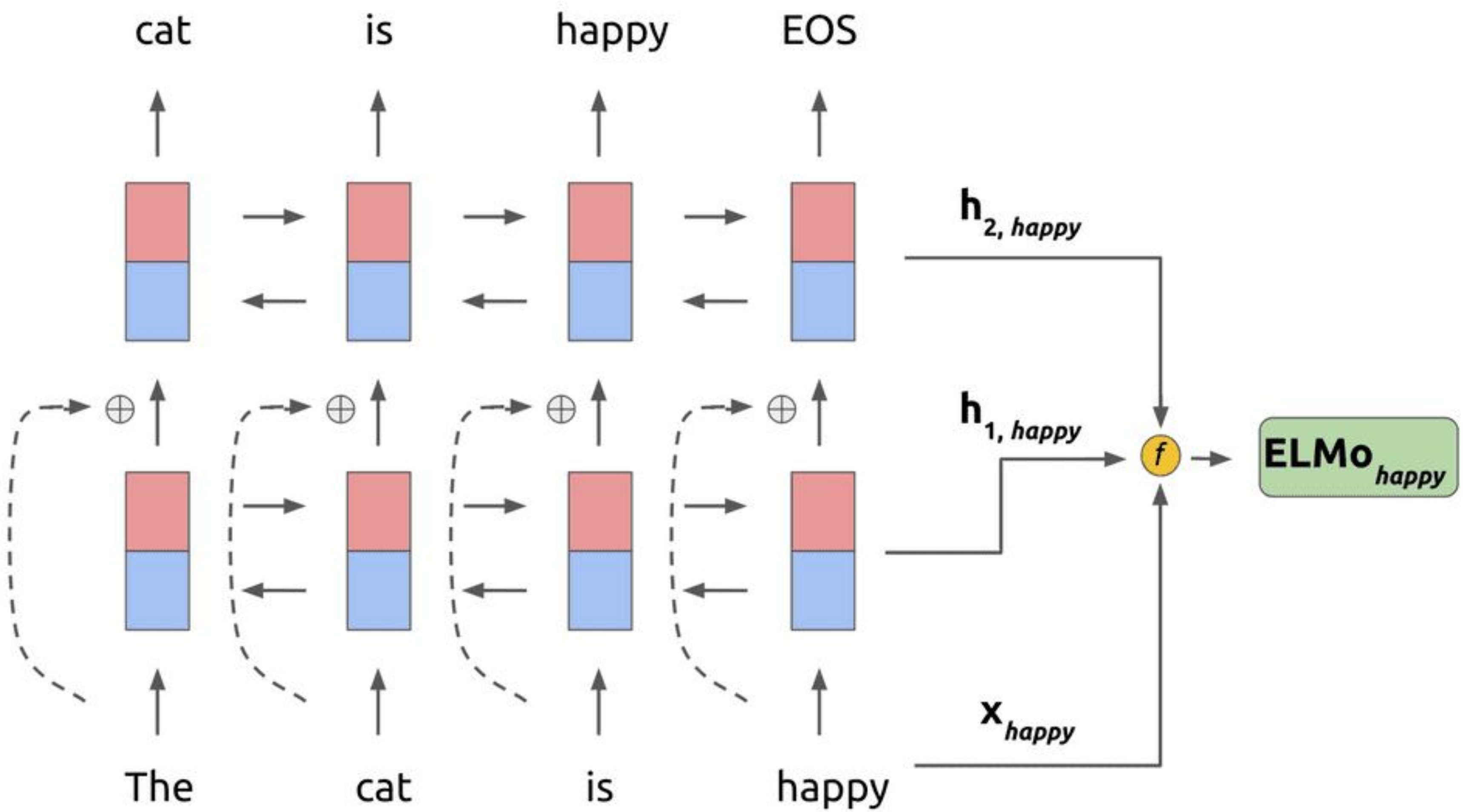
(Static) Embeddings



(Contextualized) Embeddings

- In 2018, ELMo revolutionized word embedding models
- By giving each word a different embedding depending on its context

(Contextualized) Embeddings



From: https://medium.com/saarthi-ai/elmo-for-contextual-word-embedding-for-text-classification-24c9693b0045

ELMo

Problems with ELMo:

- It's a shallow model (2 hidden layers)
- Which uses an RNN

BERT

- BERT uses transformer blocks instead of RNN layers
- It uses a much deeper network (either 12 or 24 layers)
- BERT is a deep bidirectional model
- It does not add embeddings as features, but instead performs pre-training and fine-tuning

→ BERT has entirely changed the way Deep Learning in NLP is conducted

BERT - training objectives

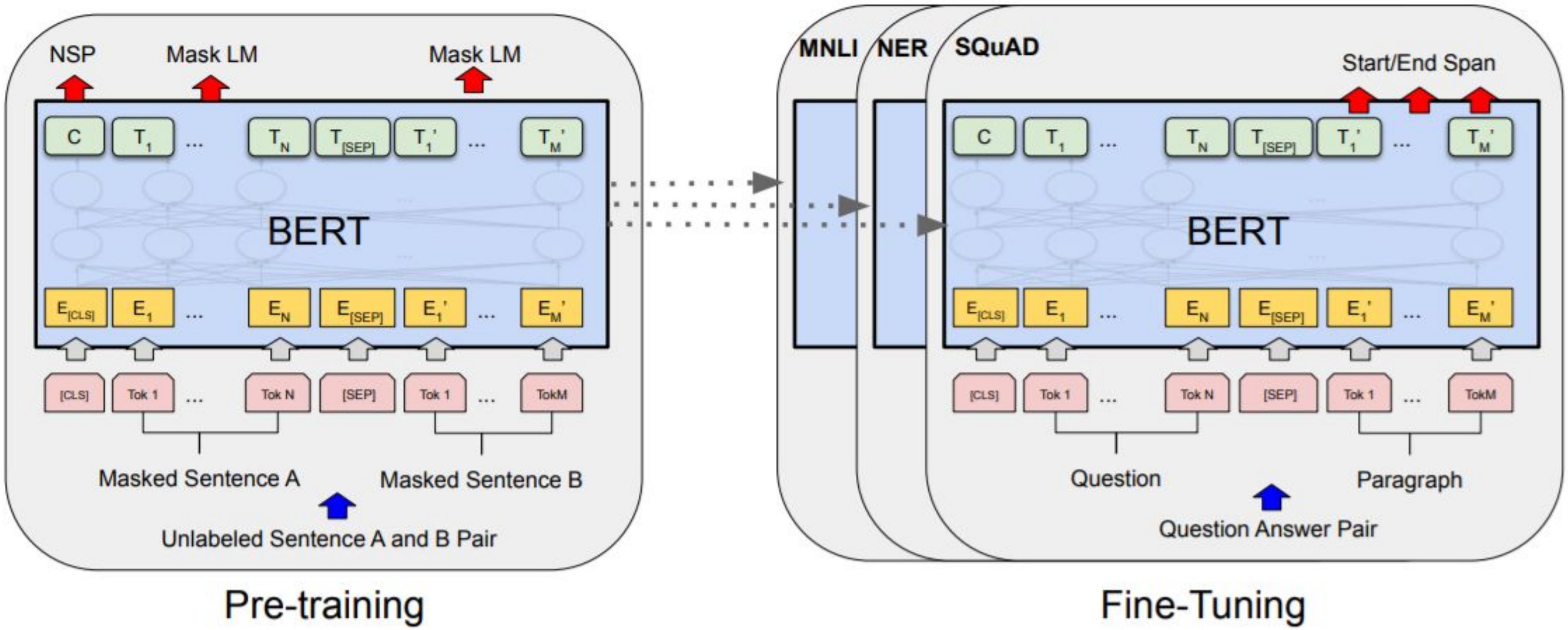
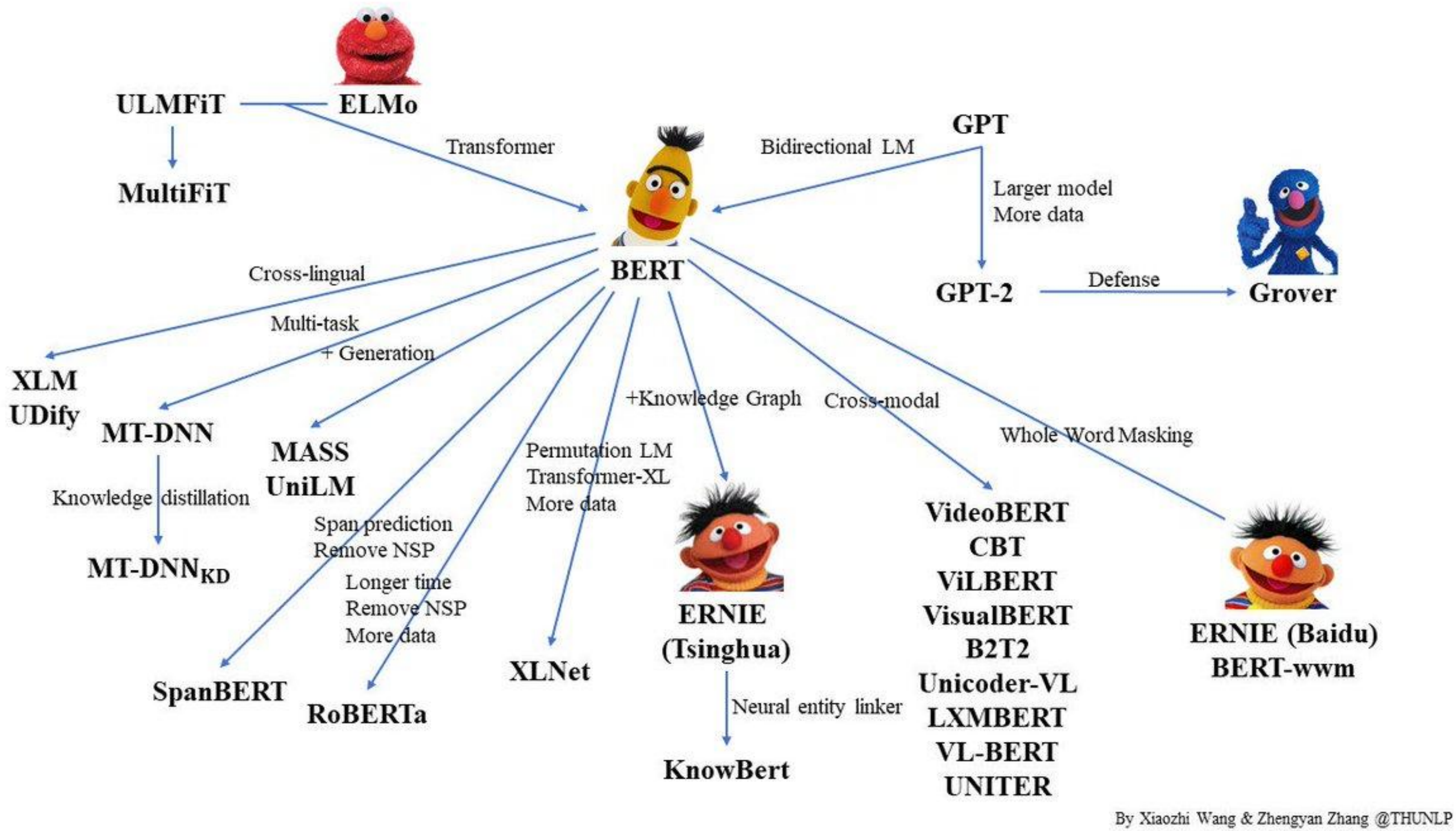


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Extensions

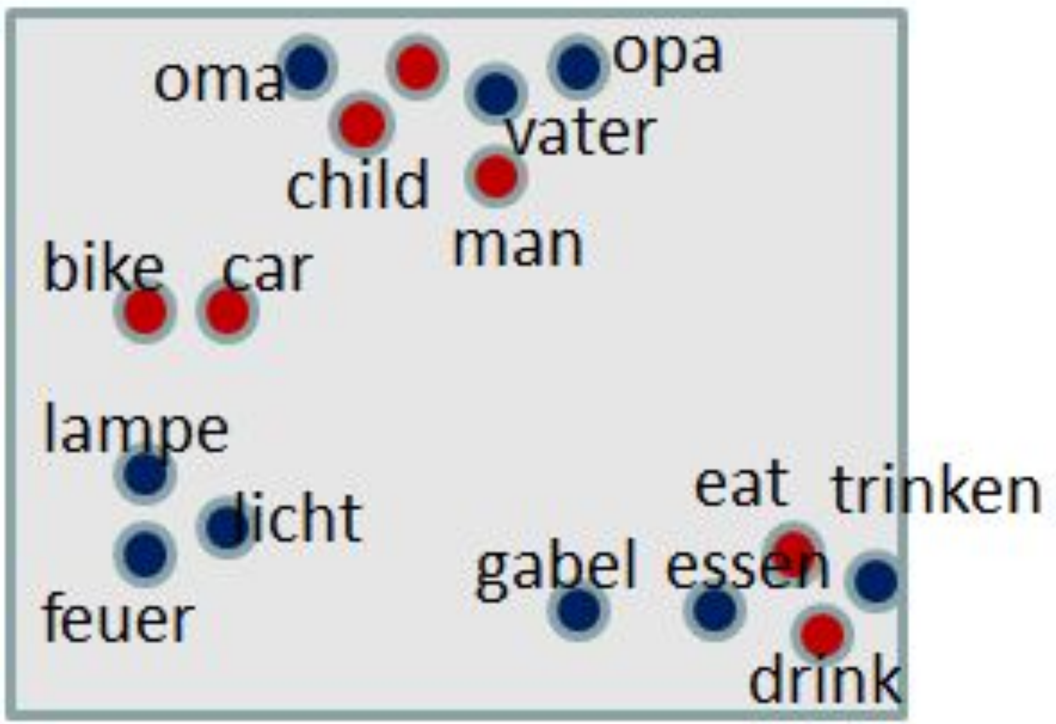
- RoBERTa (2019-07):
  - Trained for longer on more data
- ALBERT (2019-09):
  - Scaling down BERT
- MBERT (multilingual BERT) trained on the concatenation of 104 languages ...
- many others ….

Extensions


ULMFiT
ELMo
MultiFiT
Transformer
GPT
Bidirectional LM
Larger model
More data
BERT
GPT-2
Defense
Grover
Cross-lingual
Multi-task
+ Generation
XLM
UDify
MT-DNN
MASS
UniLM
Knowledge distillation
MT-DNNKD
Span prediction
Remove NSP
Longer time
Remove NSP
More data
Permutation LM
Transformer-XL
More data
+Knowledge Graph
Cross-modal
Whole Word Masking
SpanBERT
RoBERTa
XLNet
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMBERT
VL-BERT
UNITER
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

Bi- and multi-lingual embeddings (aka cross-lingual embeddings)

Similar **words** (or **sentences**) across two or more languages should be close in vector space



Cross-lingual Embeddings

Three approaches:

- **Offline methods** (e.g., Artetxe et al.)
  - Compute independent embeddings in each language, use dictionary to map in cross-lingual space
- **Joint methods** (e.g. LASER - LASER computes sentence embedding)
  - Directly leverage bilingual data at train time
- *Silly methods* (e.g. MBERT)
  - Concatenate all data and train on the concatenation

Agenda

This talk has three parts

- Contextualized Embeddings (very briefly)
- **Referenced-based Evaluation with BERT**
- Reference-free Evaluation with BERT

Traditional approach

m(y,y*), where

- y* is a human reference
- y is system prediction
- m is a "metric" based on n-gram overlaps
  - e.g. ROUGE or BLEU

Failures of ‘hard’ metrics

EN (x): „Who died two days before, and now had found \\ An unknown barren beach for burial ground“

DE-true (y*): „Vorgestern starben; dieser <u>fand</u> im Bette; \\ Des fremden Sands die letzte Ruhestätte“

DE-pred (y): „Der vor zwei Tagen starb; und nun <u>fand</u> \\ Einen unbekannten öden Strand als Grabesstätte“

Use ‘soft’ metrics instead

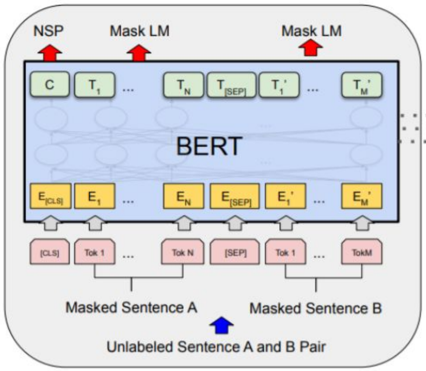
- Cannot account for lexical variation / (true) paraphrasing
- → Better approach:
    - Use word embeddings
    - Or Sentence Embeddings

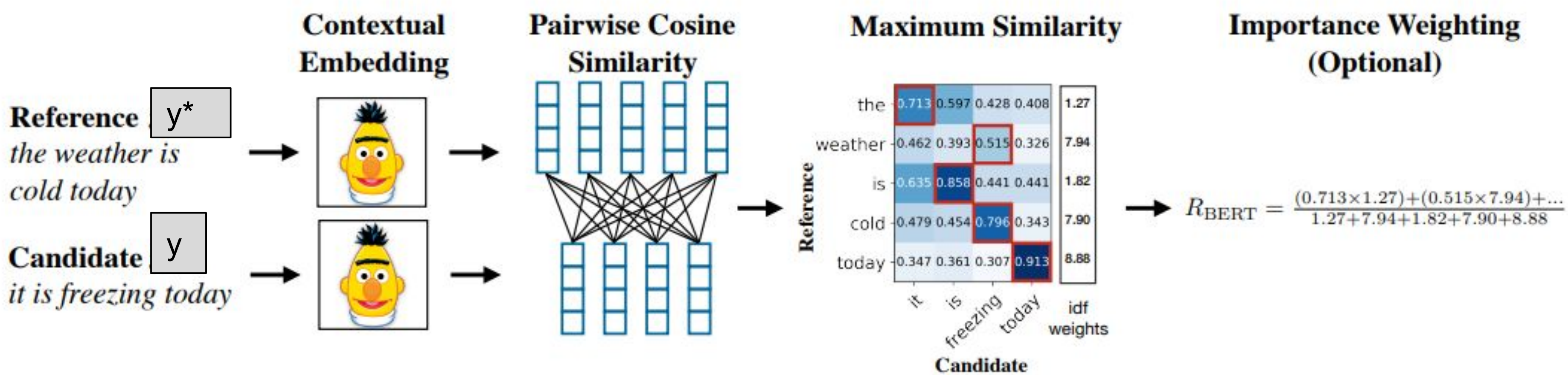
Use ‘soft’ metrics instead

TECHNISCHE UNIVERSITÄT DARMSTADT

- Cannot account for lexical variation / (true) paraphrasing
- → Better approach:
    - Use word embeddings
        - **Better than static word embeddings are contextual word embeddings**
    - Or Sentence Embeddings

Contextualized Embeddings for Evaluation





Zhang et al., BERTScore, ICLR 2020

Contextualized Embeddings for Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

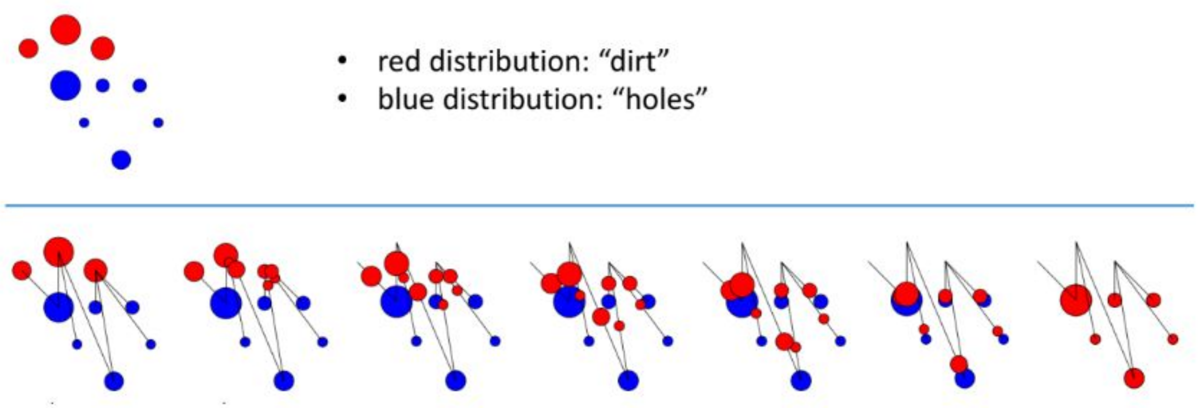
We proposed to compare two sets of contextualized embeddings with so-called **Earth Mover Distance**

- measures the amount of “work to be done” to transform one distribution into another
- Zhao et al., ***MoverScore:*** *Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance*, EMNLP 2019

Earth Mover Distance

TECHNISCHE UNIVERSITÄT DARMSTADT

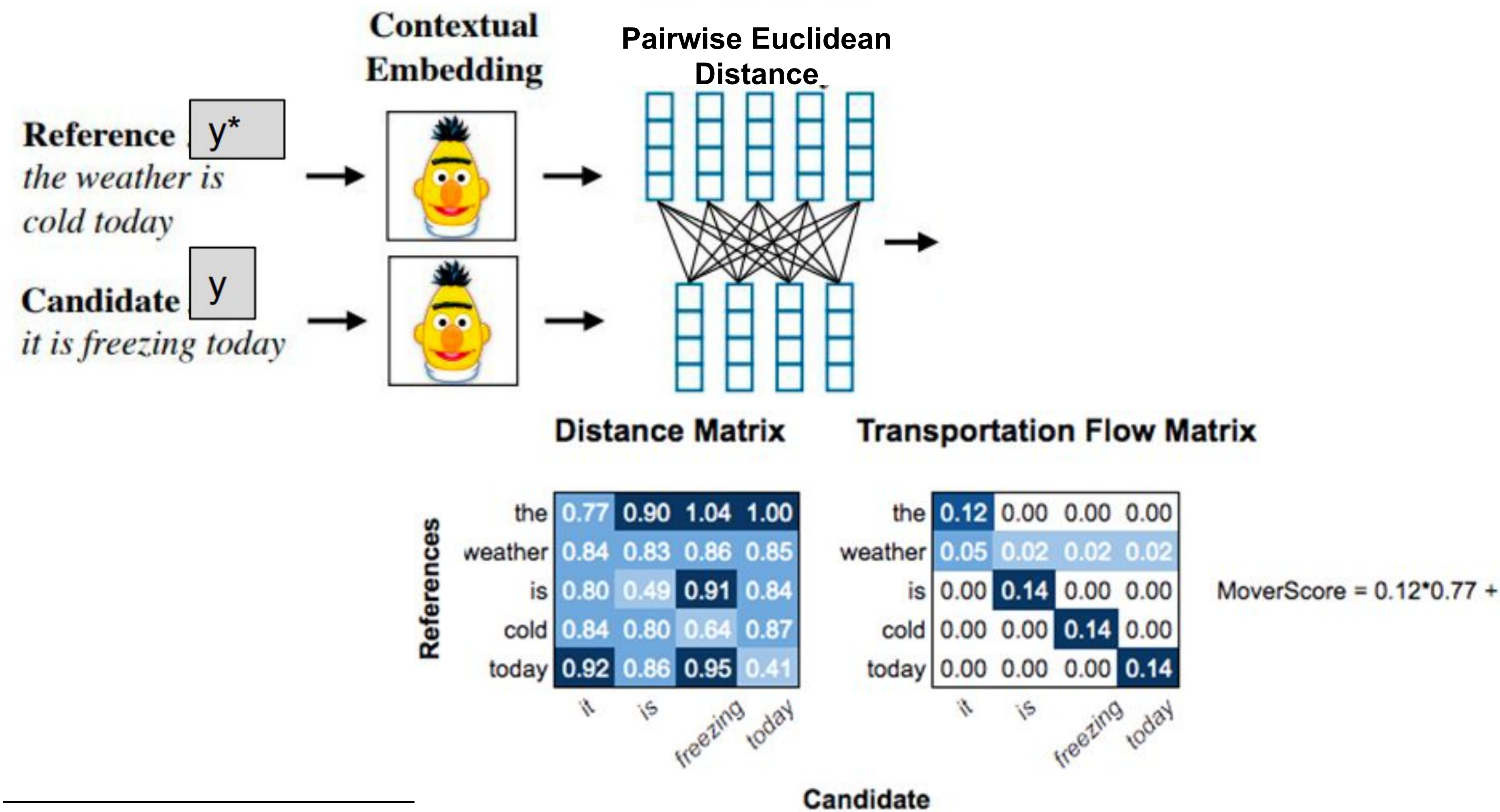
- The EMD between two distributions is proportional to the minimum amount of **work** required to convert one distribution into the other.
- The cost/work of moving the “dirt” depends on the weight/amount of “dirt” and the distance it needs to cover.



Example 2

From: https://towardsdatascience.com/earth-movers-distance-68fff0363ef2

MoverScore


Contextual Embedding
Pairwise Euclidean Distance
Reference y*
the weather is cold today
Candidate y
it is freezing today
Distance Matrix
Transportation Flow Matrix
References
the 0.77 0.90 1.04 1.00
weather 0.84 0.83 0.86 0.85
is 0.80 0.49 0.91 0.84
cold 0.84 0.80 0.64 0.87
today 0.92 0.86 0.95 0.41
it is freezing today
the 0.12 0.00 0.00 0.00
weather 0.05 0.02 0.02 0.02
is 0.00 0.14 0.00 0.00
cold 0.00 0.00 0.14 0.00
today 0.00 0.00 0.00 0.14
it is freezing today
Candidate
MoverScore = 0.12*0.77 +

BERTScore vs. MoverScore

- BERTScore uses **heuristic / greedy alignments** between words
- MoverScore computes **optimal alignment** by solving an optimization problem

Agenda

This talk has three parts

- Contextualized Embeddings (very briefly)
- Referenced-based Evaluation with BERT
- **Reference-free Evaluation with BERT**

Extensions

Most metrics today still use human references y*

- Costly to obtain
- Evaluation is limited to the parallel data available

Can we get rid of human references y* and instead only use (x,y) for evaluating text generation?

- where x is the source sentence(s)

RFEval

In other words, we aim for metrics

m(x,y)

instead of

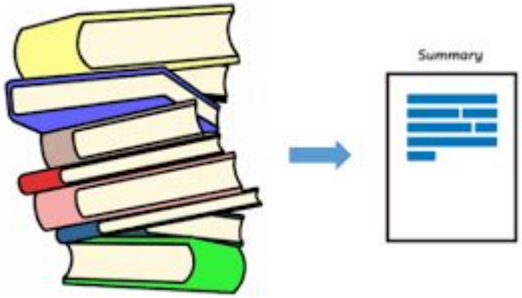
m(y*,y)

Tasks

Evaluation of:

- Machine translation

- Summarization



How to do RFEval

1. Directly assess the quality/similarity of (x,y)
2. Create a pseudo-reference y** and evaluate (y**,y)

How to do RFEval

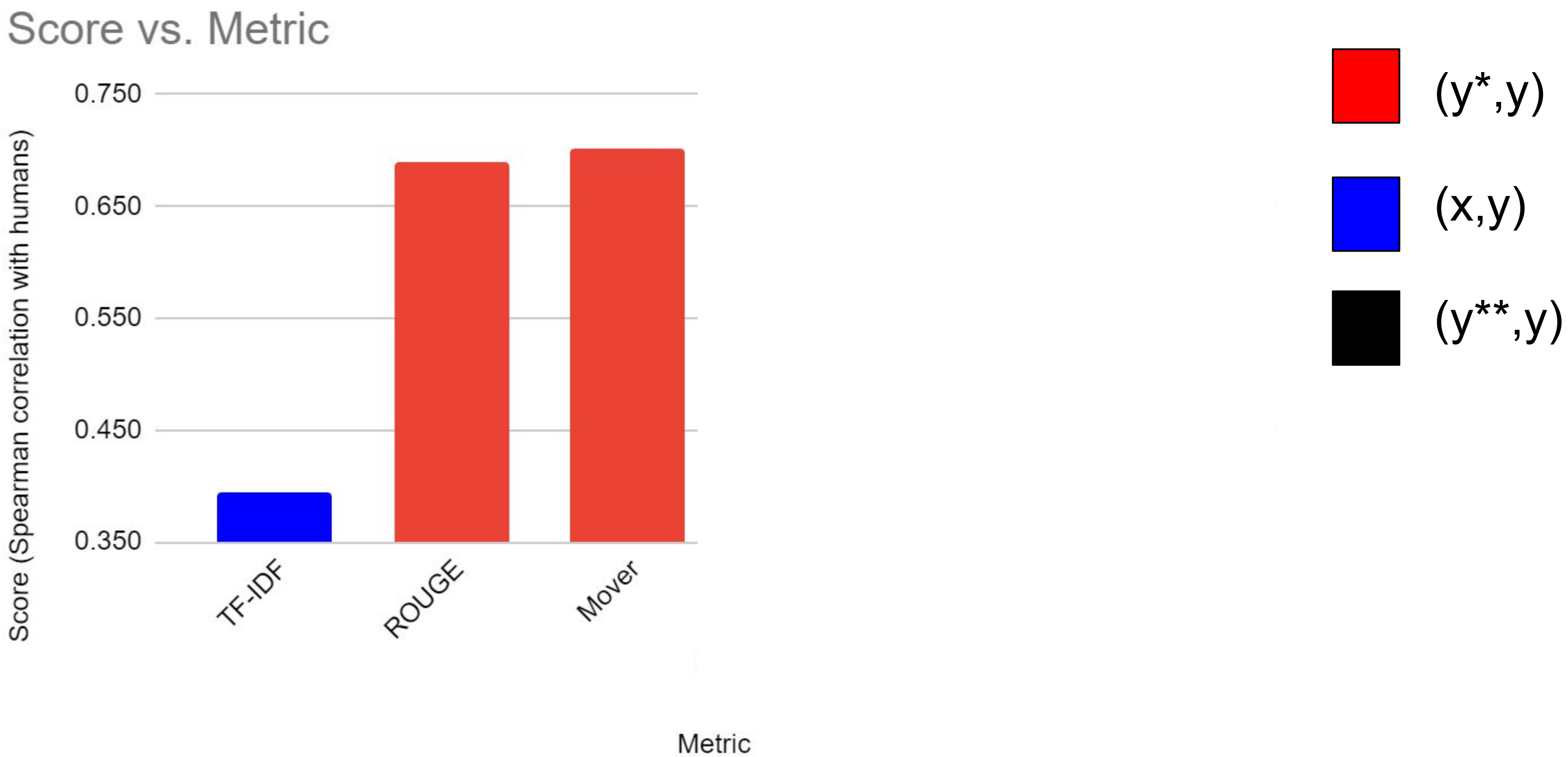
TECHNISCHE UNIVERSITÄT DARMSTADT

1. Directly assess the quality/similarity of (x,y)
   - In MT: **cross-lingual space** (LASER, MBERT, XUSE, ...)
   - In Summarization: mono-lingual space (but with enormous **length differences**)

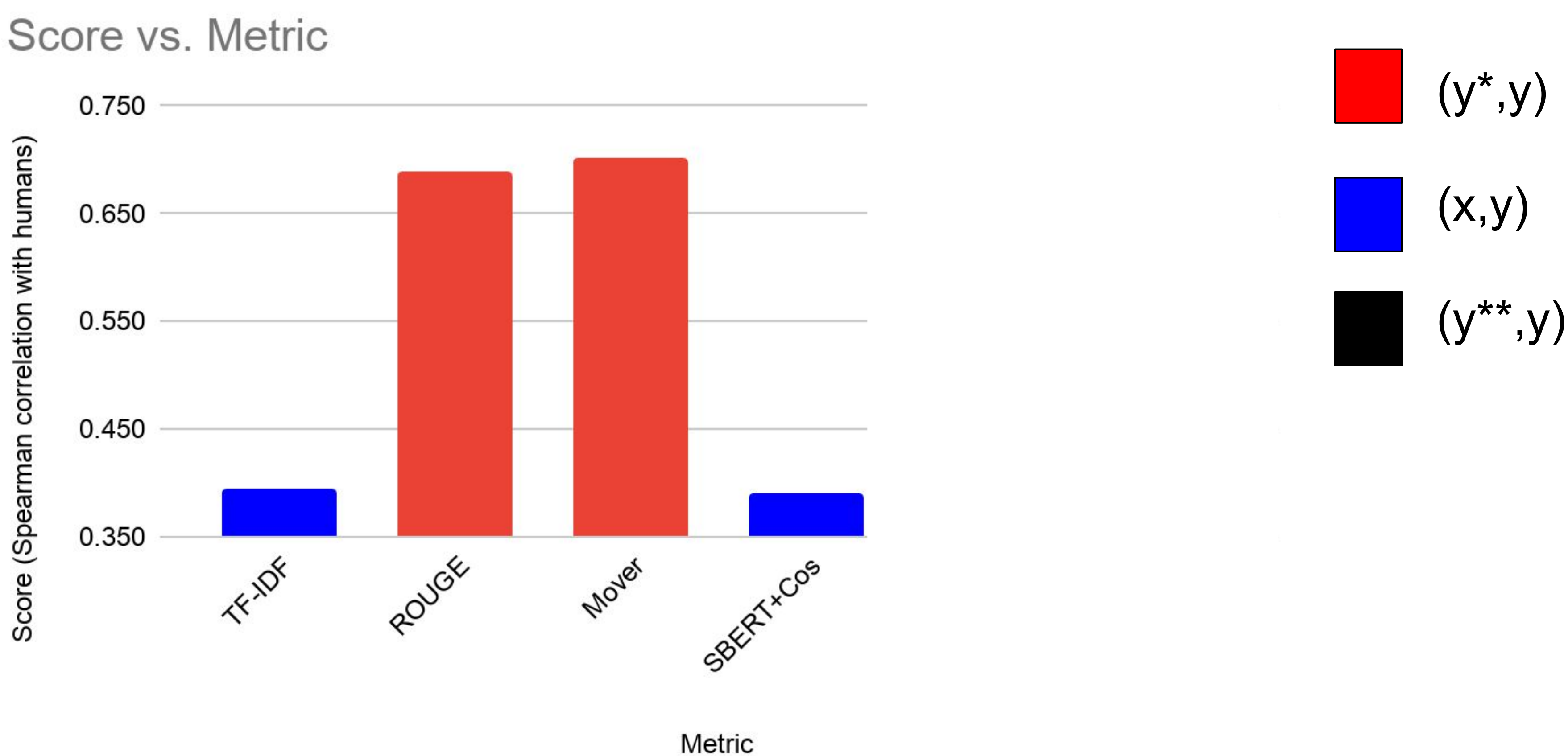
How to do RFEval

2. Create a pseudo-reference y** and evaluate (y**,y)
    - In MT:
        - Google translate
        - Unsupervised (N)MT
    - In Summarization:
        - Keep important sentences in source documents

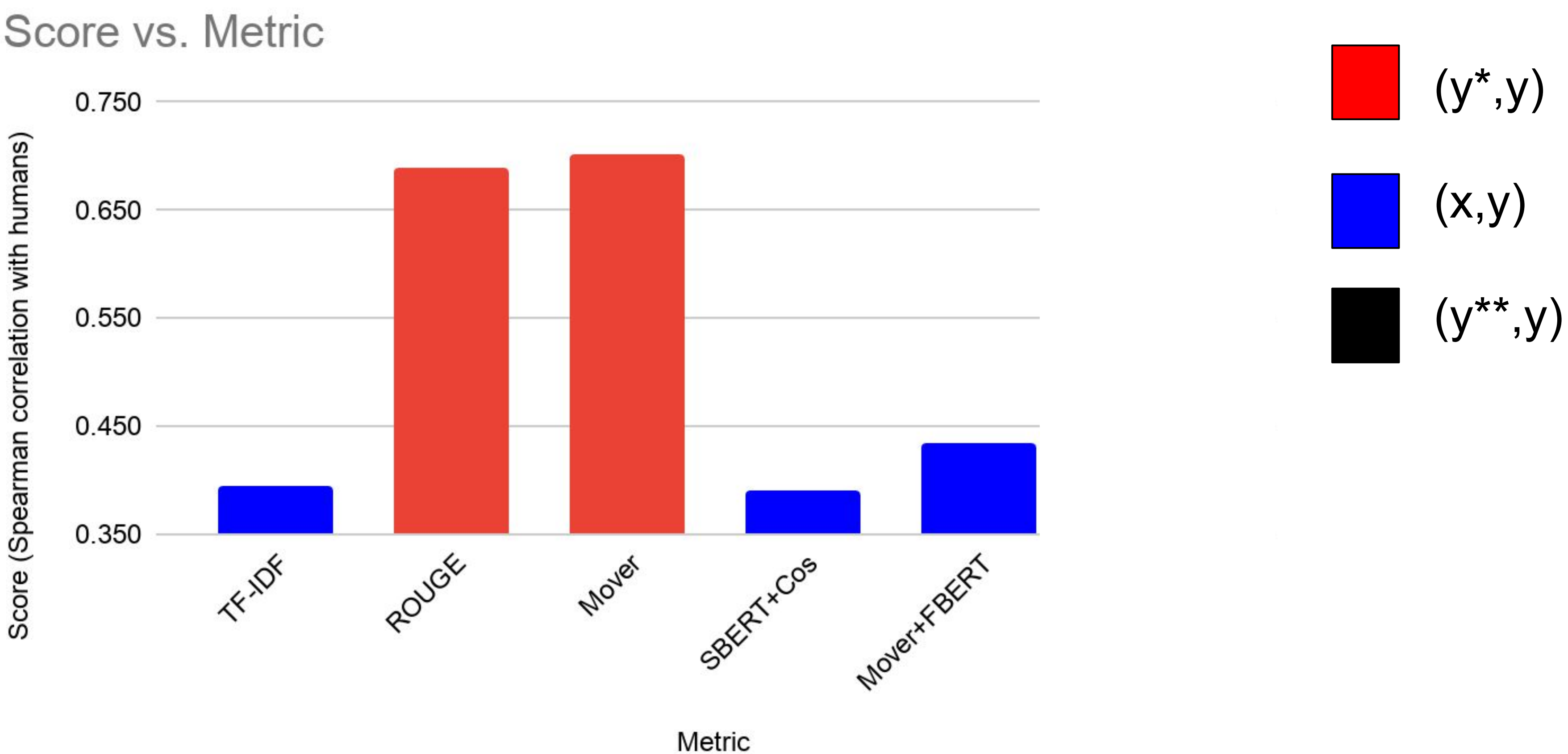
Results for Summarization



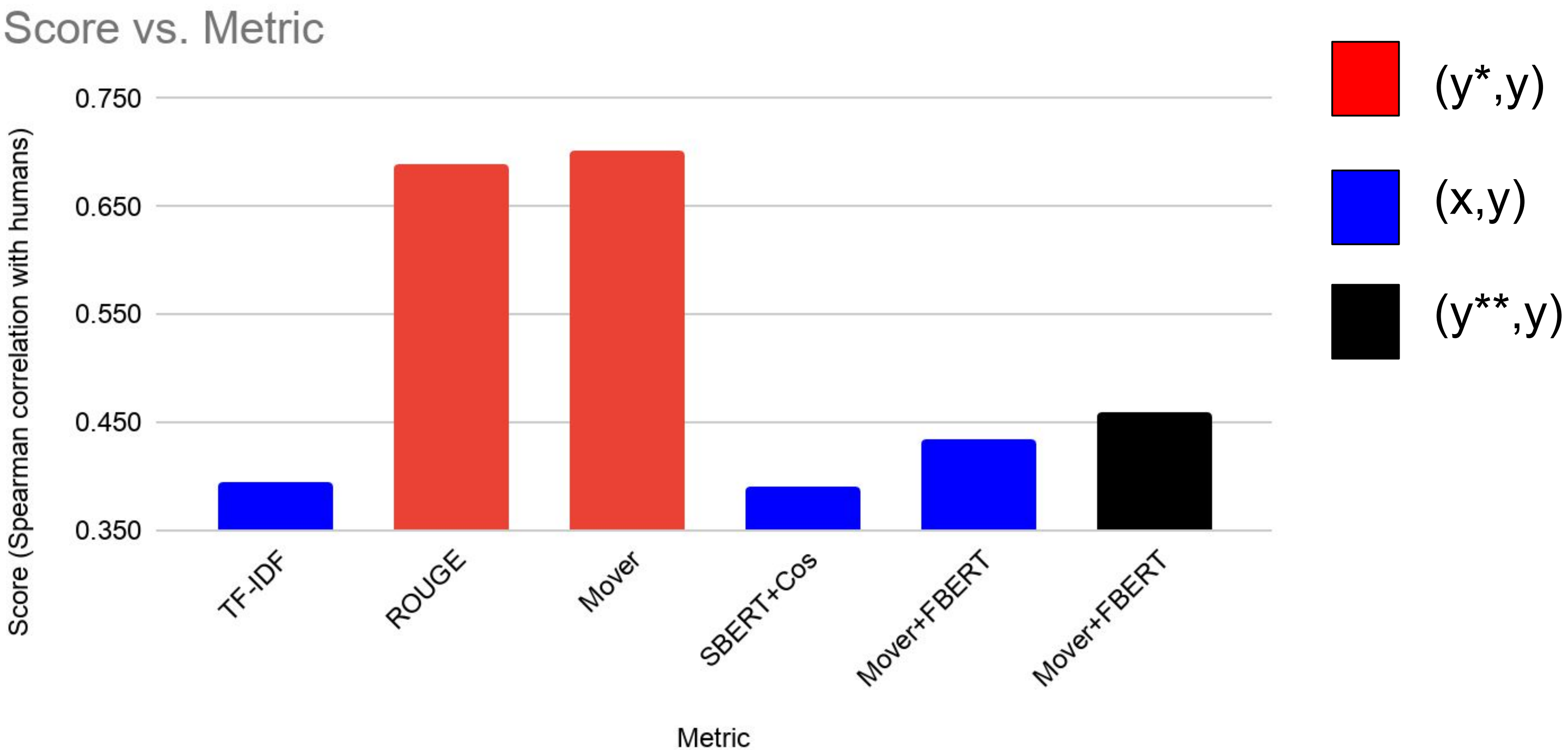
Results for Summarization



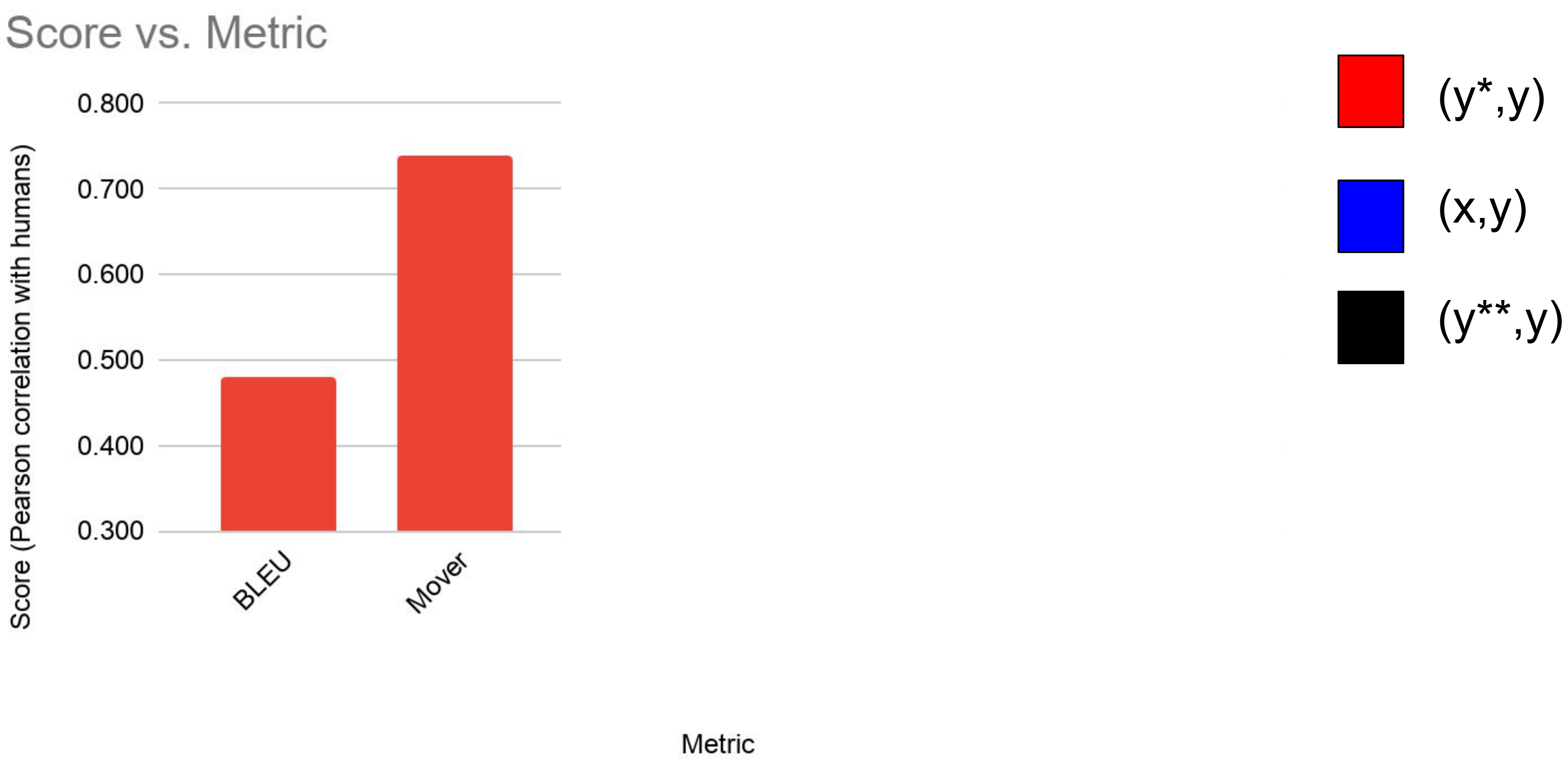
Results for Summarization



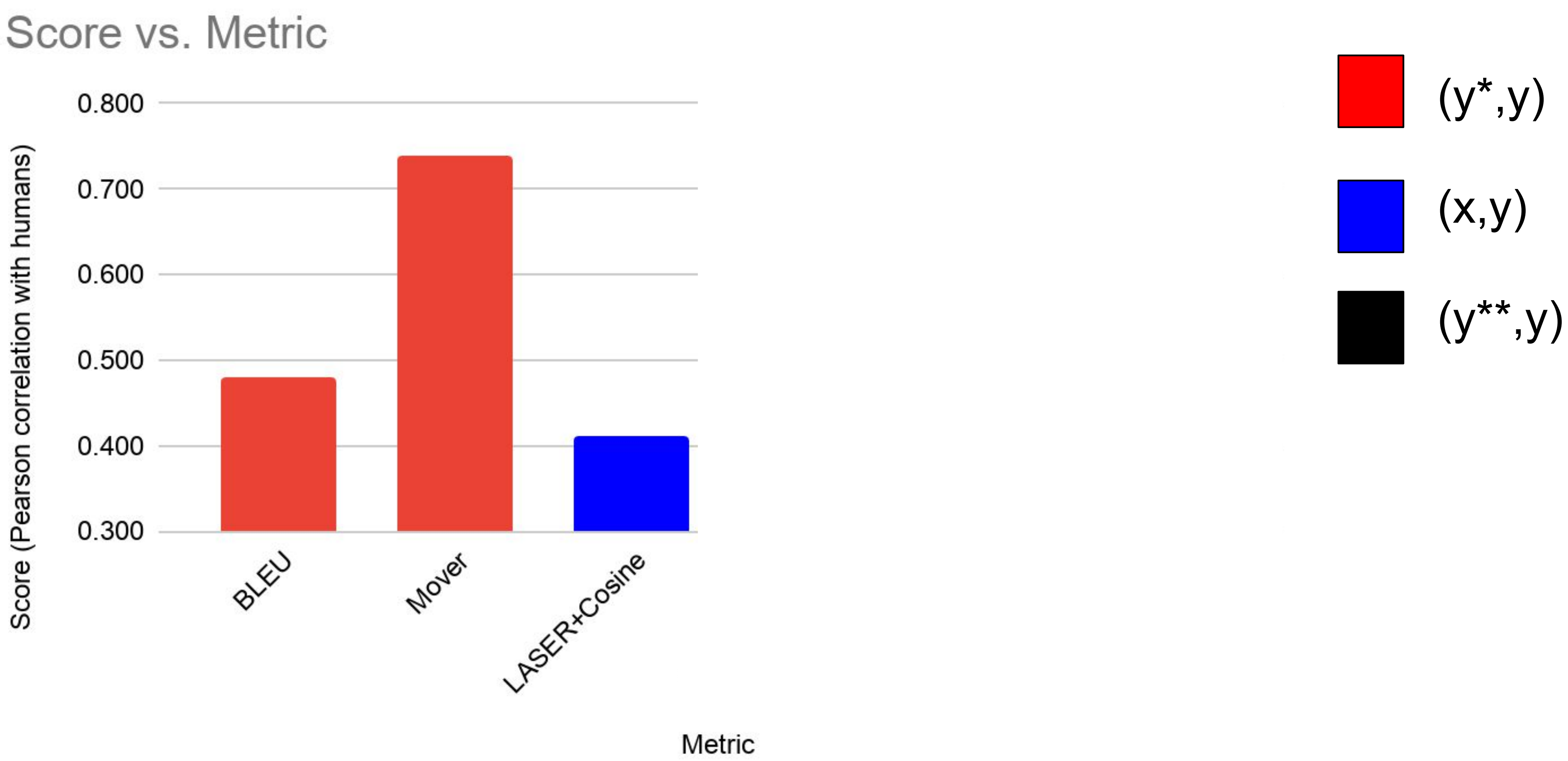
Results for Summarization



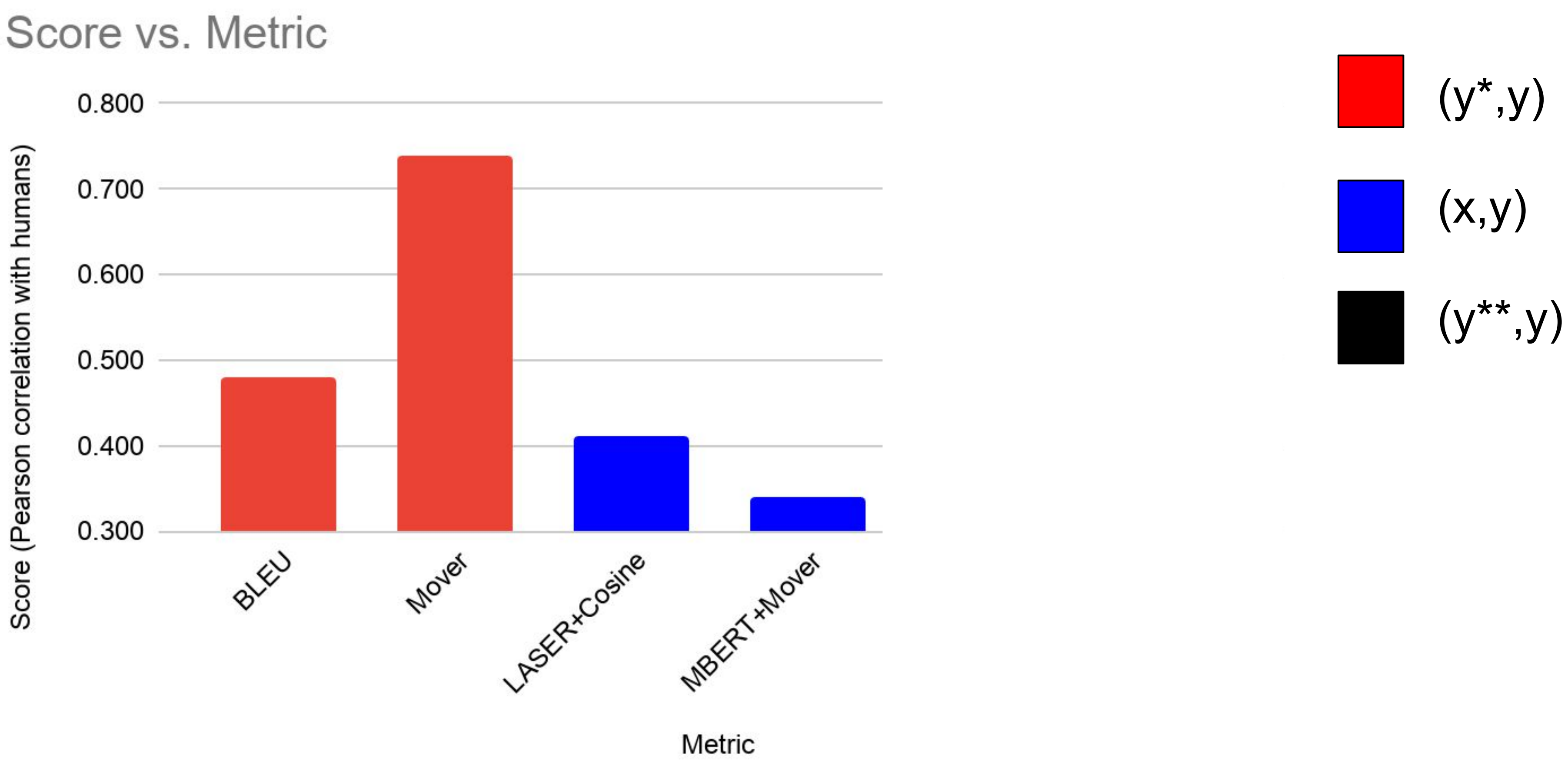
Results for MT



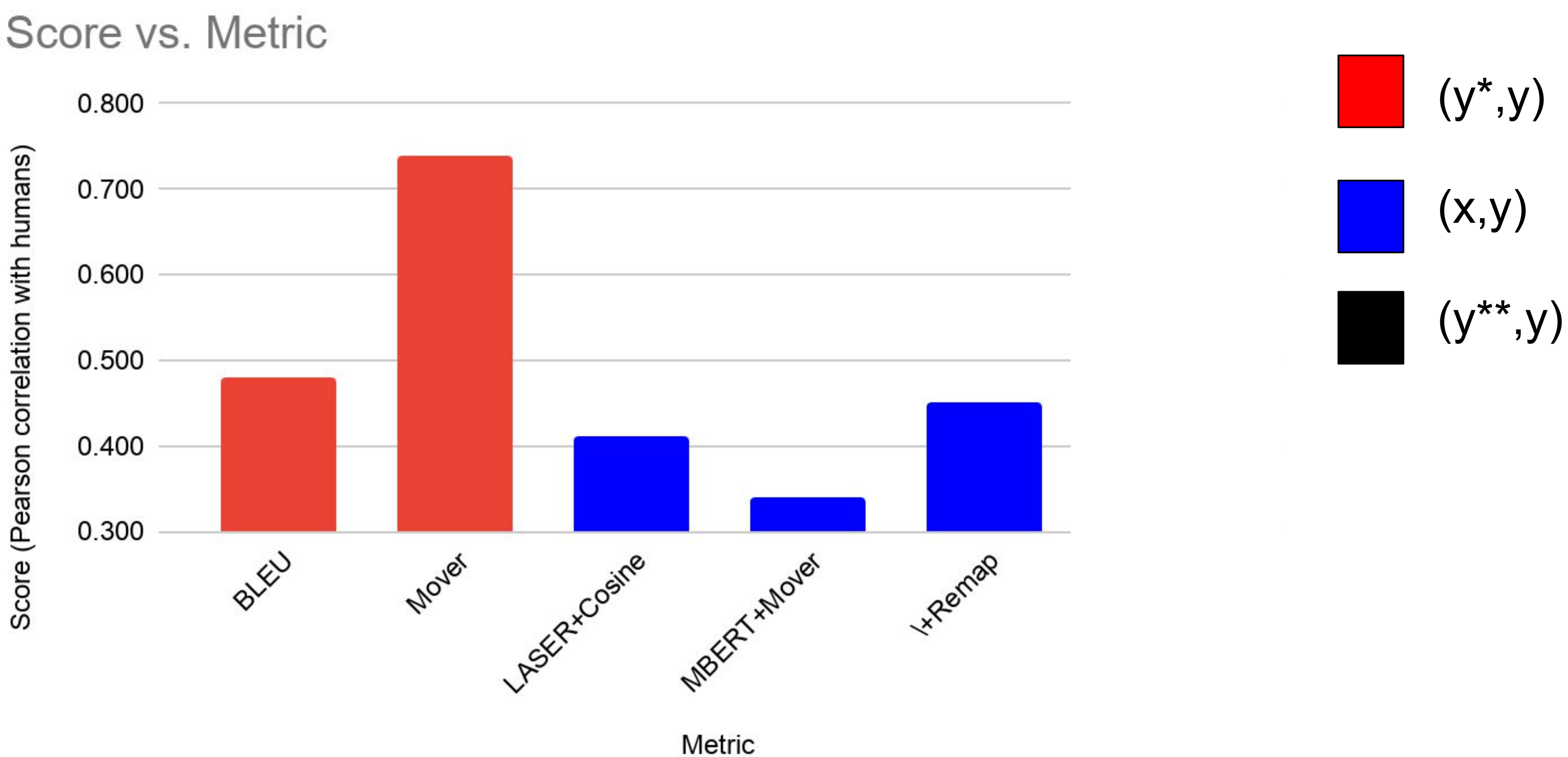
Results for MT



Results for MT

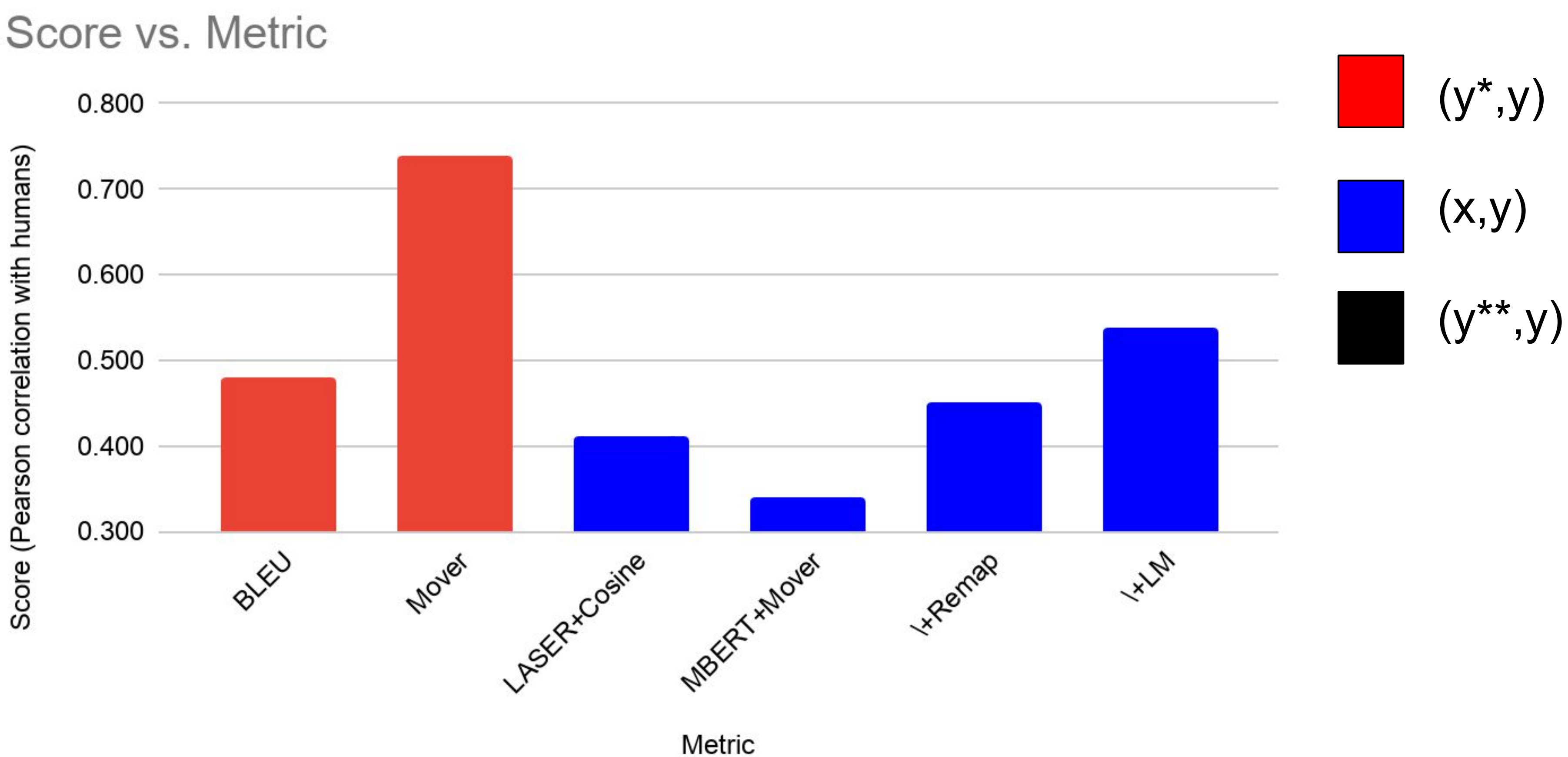


Results for MT

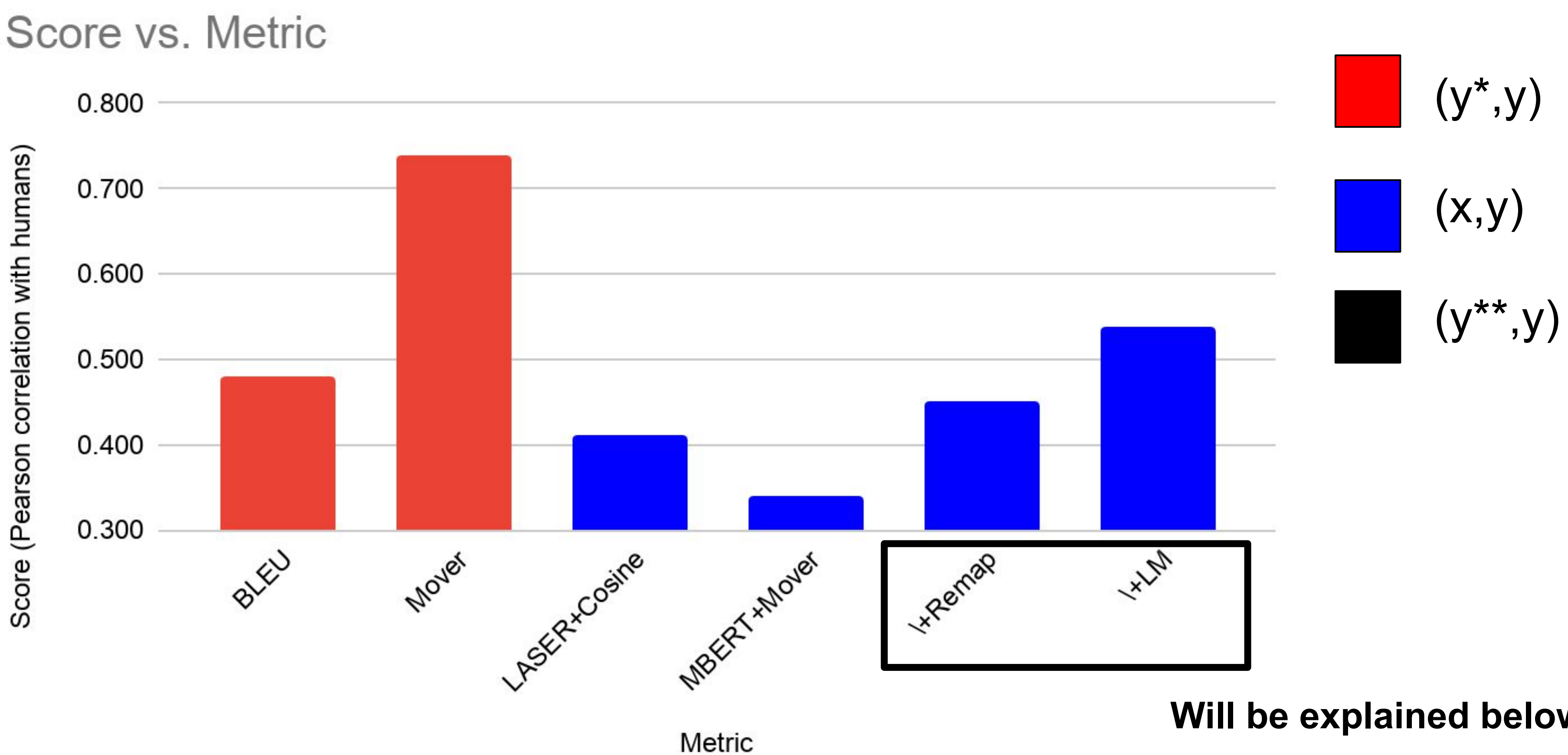


Results for MT

TECHNISCHE UNIVERSITÄT DARMSTADT



Results for MT



Will be explained below

Why do cross-lingual encoders fail for MT?

TECHNISCHE UNIVERSITÄT DARMSTADT

Suspicion:

- (1) MBERT may be ill-aligned (maybe LASER not so much)
- (2) MT systems often produce very literal translation ("translationese")
  - Krankheitsfall: Wann bezahlt der veranstalter
  - Disease case: when paying the organiser

  How do cross-lingual encoders deal with translationese?

(1) Ill-alignment

To address that MBERT hasn't seen any parallel data, we **Remap** the MBERT space using a bit of parallel data:

- We acquire bilingual data from EuroParl
- And learn a linear projection matrix, similar to the approach of Mikolov (2013)

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\boldsymbol{X}_\ell - \boldsymbol{X}_k\|_2.$$

X_l and X_k contain corresponding vectors for languages l,k

(2) Translationese

TECHNISCHE UNIVERSITÄT DARMSTADT

We fool/probe them using the following:

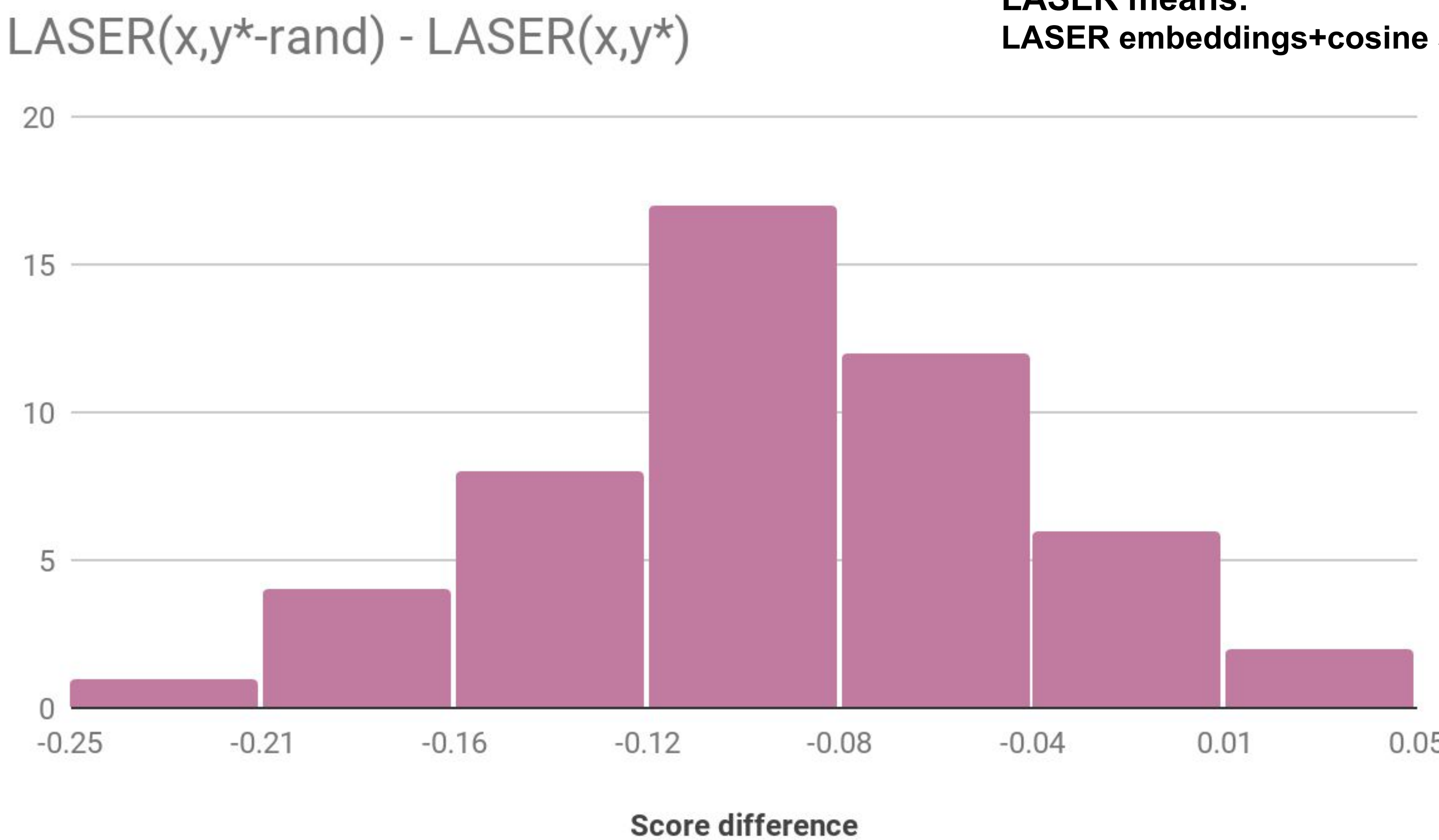
- Random Shuffle y*
- Expert reordering y* (to match word order of x)
- Expert Word-by-Word translation of x

Robustness of Cross-lingual representations

x	Dieser von Langsamkeit geprägte Lebensstil scheint aber ein Patentrezept für ein hohes Alter zu sein.
y*	however, this slow pace of life seems to be the key to a long life.
y*-rand	to pace slow seems be the this life. life to a key however, of long
y*-reord	this slow pace of life seems however the key to a long life to be.
W2W	this of slow pace characterized life style seems however a patent recipee for a high age to be.

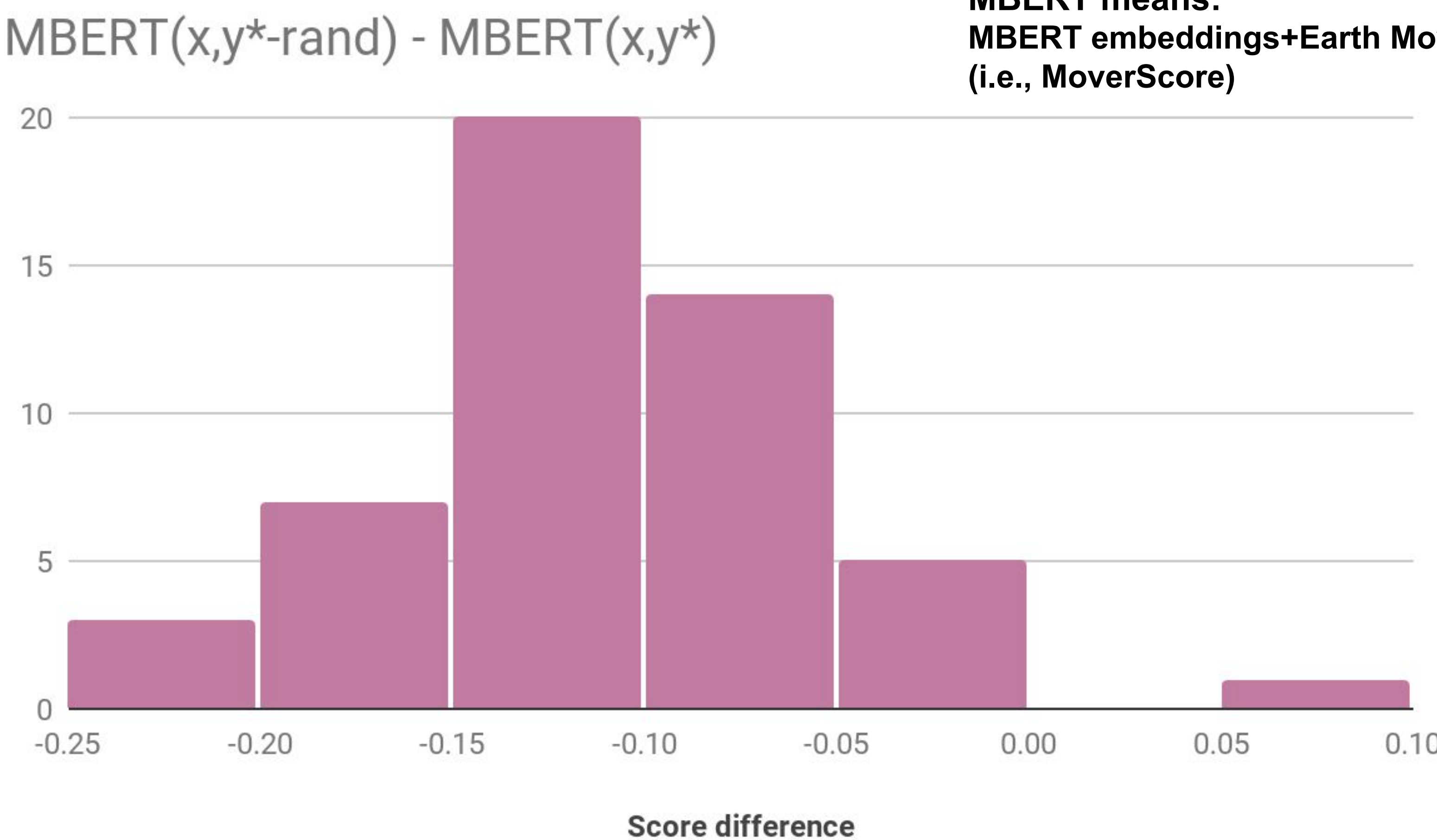
Random shuffle

LASER means:
LASER embeddings+cosine similarity

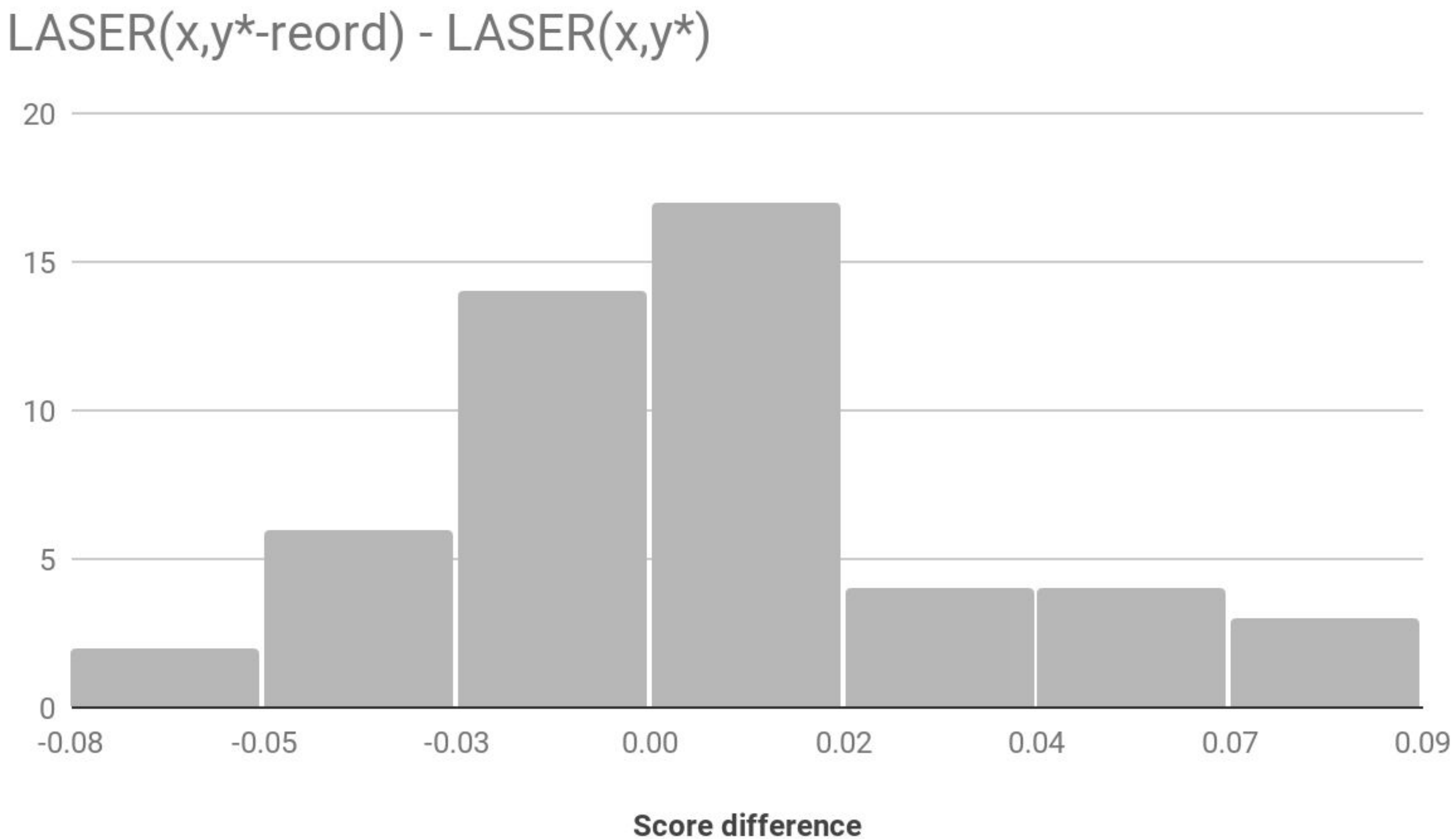


Random shuffle

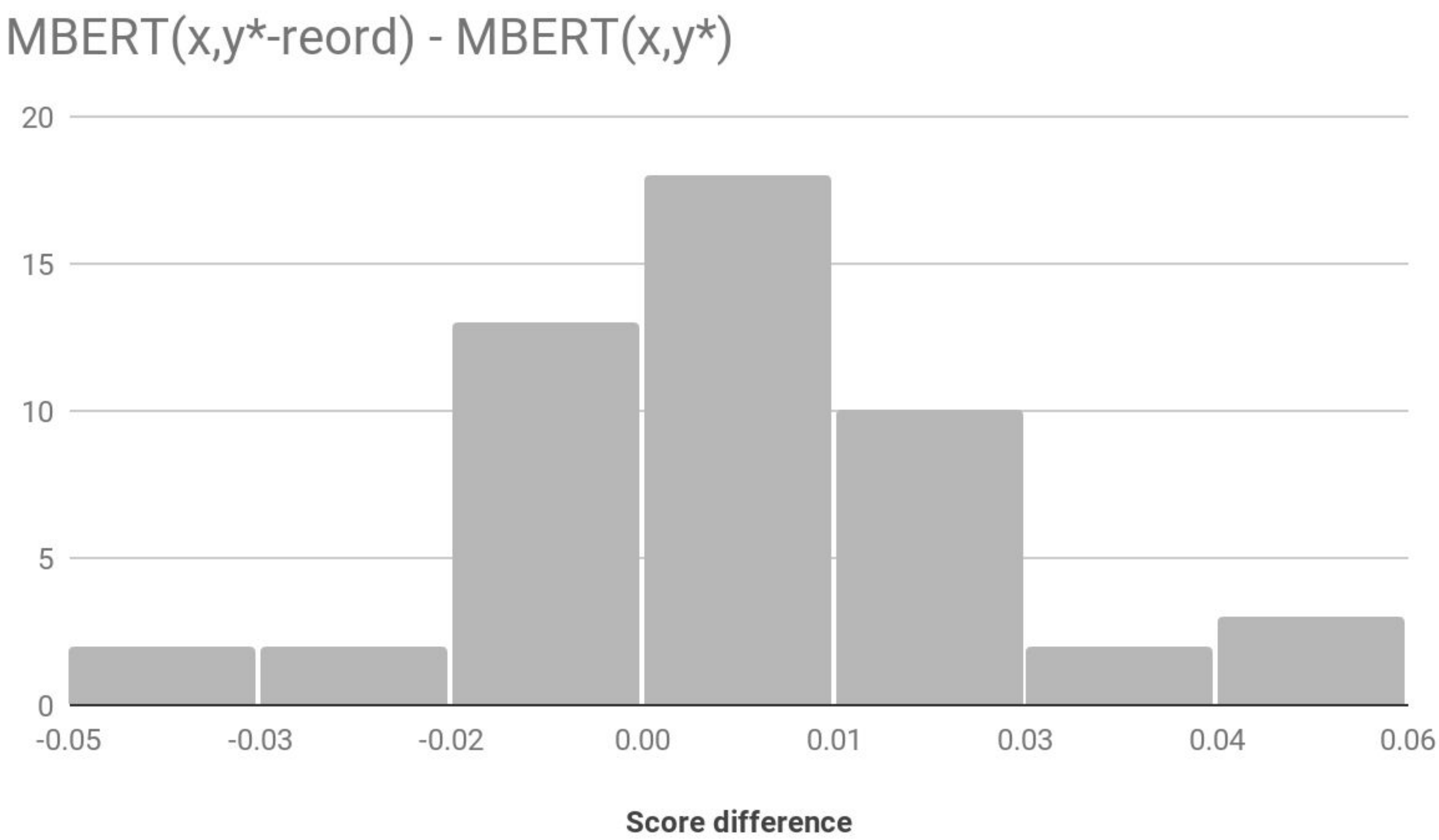
MBERT means:
MBERT embeddings+Earth Mover Dist
(i.e., MoverScore)



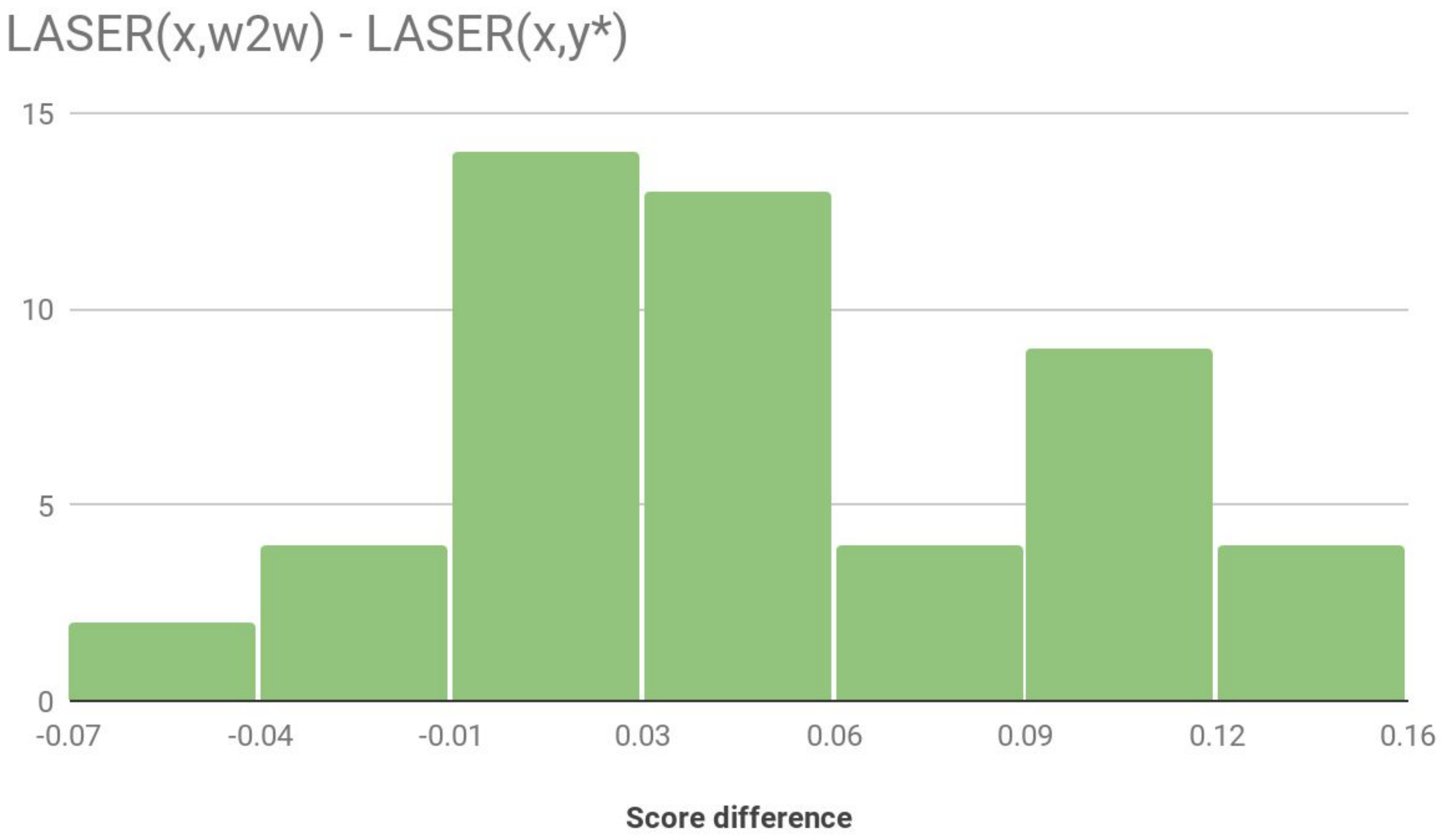
Expert reordered



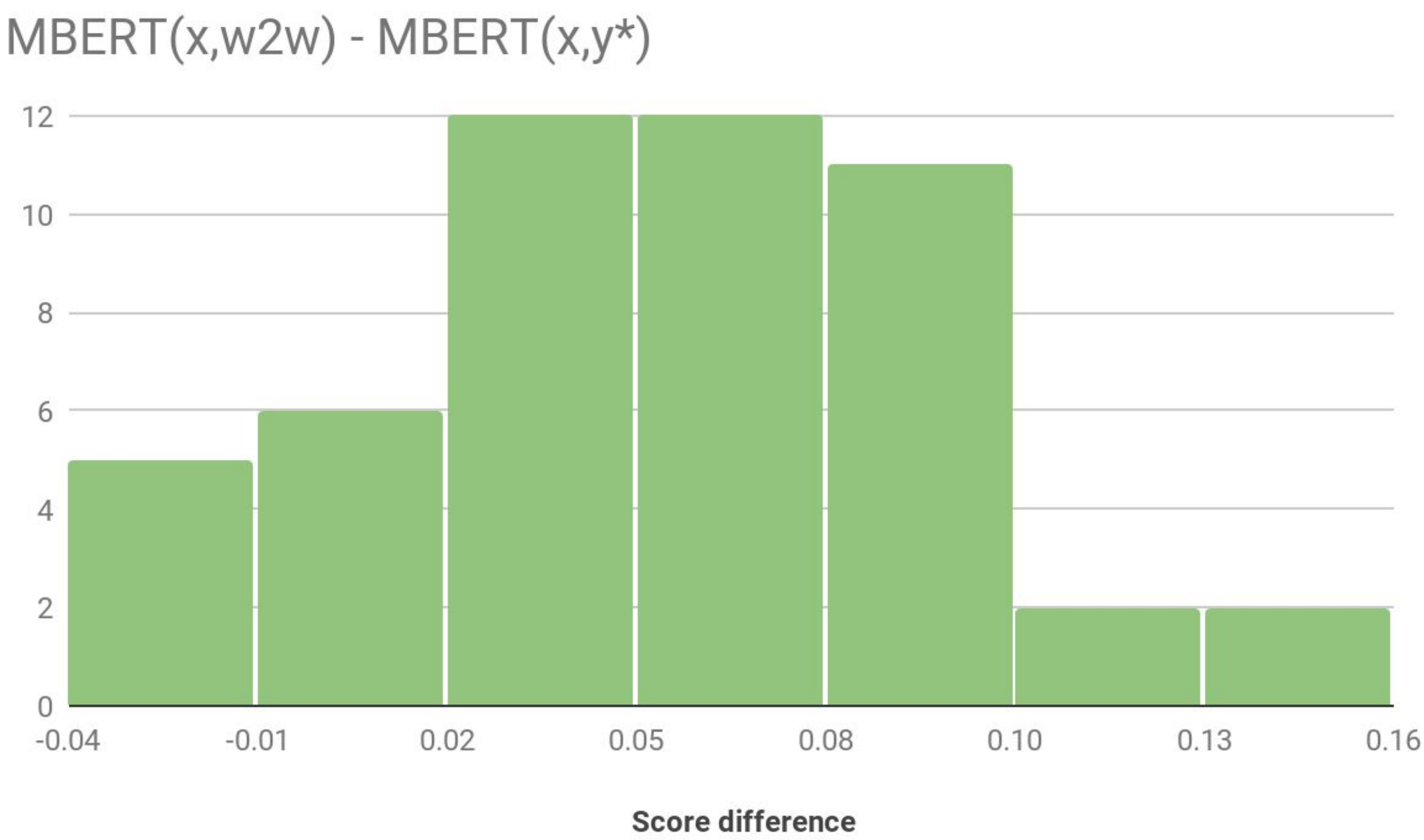
Expert reordered



W2W


LASER(x,w2w) - LASER(x,y*)
15
10
5
0
-0.07
-0.04
-0.01
0.03
0.06
0.09
0.12
0.16
Score difference

W2W



MBERT(x,w2w) - MBERT(x,y*)
12
10
8
6
4
2
0
-0.04
-0.01
0.02
0.05
0.08
0.10
0.13
0.16
Score difference

(2) Translationese

- To fix the translationese issue, we add a language model **LM** to our cross-lingual embeddings:

$$m(x,y) = 0.9 * ce(x,y) + 0.1 * LM(y)$$

Conclusion

- Rapid advances due to Eval. Metrics based on BERT
- In the reference-free context still a considerable gap
- We exposed (severe) deficits of current cross-lingual sentence encoders / metrics
  - While not BOW models
  - They are indifferent between correct and source language word order
  - They like W2W **("translationese")** - which is a severe problem for MT evaluation metrics

THÄNK$!

References

- Zhao, Peyrard, Liu, Gao, Meyer, Eger. *MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance*. EMNLP 2019
- **Gao, Zhao, Eger. *SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization.* ACL 2020**
- **Zhao, Glavas, Peyrard, Gao, West, Eger. *On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation.* ACL 2020**