Metrics in NLP

Jai Javeria, COV 884

Monday, 18 April 2022

Index

- Motivation
- Some Old Metrics
  - ~~Monday~~ BLEUs
  - ROUGE
  - METEOR
- Some New Metrics
  - MoverScore
  - BERTScore
  - Student Reviews
  - MoverScore vs BERTScore

Motivation

- NLP is a large field with various tasks, each with various different metrics
  - Machine Translation (BLEU, RUSE)
  - Image Captioning (SPICE, LEIC)
  - Summarization (ROUGE, s^3_{best})
  - Question Answering (F1 Score)
  - Generation (Perplexity)
- One can look at some metrics that can be used over multiple tasks
- Many evaluation tasks require to find optimal sentence similarity
  - They have a reference sentence. They check if the system generated sentence is similar to the previous one. If yes, then a higher score is given
  - Prevalent across various tasks: MT, Image Captioning, Summarization, QA etc

BLEU

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.

Any guesses on the number of citations of BLEU?

BLEU

- Number of citations as of today: 18006
- Compute the n grams and calculate the precision $\text{Exact-P}_n = \frac{\sum_{w \in S_{\hat{x}}^n} \mathbb{I}[w \in S_x^n]}{|S_{\hat{x}}^n|}$
- Reference Sentence: I ate three Hazelnuts
  - Candidate 1: three three three three
  - Exact Precision would give this high score.
  - Update: Each n-gram can be match at most once.
  - Candidate 2: I ate
  - The updated metric still would give this a high score
  - Update: Add a brevity penalty
- the number of exact matches is accumulated for all reference-candidate pairs in the corpus and divided by the total number of n-grams in all candidate sentences

Referenced from BERTScore Related Work Section and this medium article

ROUGE

Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text summarization branches out*. 2004.

ROUGE

- A recall based metric that measures the degree of lexical overlap between system output and a set of reference summaries.
- Has multiple variants
  - ROUGE-N: computes similarity of N grams
  - ROUGE-L: measures the longest common subsequence between the output and reference
  - ROUGE-S: uses skip grams. e.g. reference: “the fox”, candidate: “the brown fox”

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

METEOR

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

METEOR

- Computes unigram precision, recall and FMean(H.M. between P and 9R).
- Can also match word stems, synonyms and paraphrases. E.g. run and running.
- Would require a stemmer, synonym lexicon and paraphrase table. Limits the number of languages on which it can be used.
- Puts a penalty if word order not followed, using concept of chunks.
- Chunk: group of adjacent unigrams in the system translation that are mapped to adjacent unigrams in the reference translation.
- Score = (1-Penalty)*Fmean

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

Towards Contextualized Embeddings

- Reference Sentence: people like foreign cars
  - Candidate 1: people like visiting foreign countries
  - Candidate 2: consumers prefer imported vehicles
  - Ngram metrics give higher score to first sentence rather than the second
- BLEU will only mildly penalize swapping of cause and effect clauses (e.g. A because B instead of B because A)
- Solution: use contextualized embeddings which can capture semantic similarity

MoverScore

Zhao, Wei, et al. "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019.

Introduction

- Uses Word Mover Distance to calculate similarity
- Let x,y be a sequence of n grams
- Embedding of a n-gram is the weighted (idf) sum over its word embeddings
- Let d be a euclidean distance between embeddings of two n-grams
- Let C be the transportation cost matrix such that $C_{ij} = d(\mathrm{x}_i^n, \mathrm{y}_j^n)$
- Let f_{x^n} and f_{y^n} be the associated weights of the sequence of n grams

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\mathsf{T}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

$$\boldsymbol{f}_{\boldsymbol{x}_i^n} = \frac{1}{Z} \sum_{k=i}^{i+n-1} \mathrm{idf}(\mathrm{x}_k)$$

where Z is a normalizing constant s.t. $\boldsymbol{f}_{\boldsymbol{x}^n}^\mathsf{T}\mathbf{1} = 1,$

〈C,F〉 denotes the sum of all matrix entries of the matrix C ⊙ F where ⊙ denotes element wise multiplication. Use linear programming to get WMD (cubic time)

How does the embedding of a token come from Model

- In models like BERT, each layer has a representation of the token given.
- Now the question comes is that how do we choose which representation to give to our metric.
- It has been shown that intermediate layers have better representation.
- In BERTScore, they fix a layer and experiment to see which layer is the best
- In MoverScore, they use an aggregation map to combine the representations
- They use power means as it is a generalization of pooling
- Let $z_{i,l}$ be the representation of the i^{th} word given by the l^{th} layer. Let p be a number, including infinity

$$\boldsymbol{h}_i^{(p)} = \left(\frac{\boldsymbol{z}_{i,1}^p + \cdots + \boldsymbol{z}_{i,L}^p}{L} \right)^{1/p} \in \mathbb{R}^d$$

where exponentiation is applied elementwise.

$$E(x_i) = \mathbf{h}_i^{(p_1)} \oplus \cdots \oplus \mathbf{h}_i^{(p_K)}$$

where $\oplus$ is vector concatenation; $\{p_1, \ldots, p_K\}$ are exponent values, and we use $K = 3$ with $p = 1, \pm\infty$ in this work.

Dimensions of MoverScore

- Granularity
  - What should be the value of n? Experiments conducted with unigram, bigram and sentence level
- Embedding Choice
  - Experiment with word2vec, ELMo and BERT
- Fine Tuning
  - The fine tune the model on tasks so that they can get better embeddings. Use 2 NLI datasets, MultiNLI and QANLI and Paraphrase dataset QQP (Quora Question Pair)
- Tasks
  - Report the efficacy of the metric on Machine Translation, Summarization , Image Captioning and Dialogue Response Generation

Results- Machine Translation

Setting	Metrics	Direct Assessment							
		cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	RUSE(*)	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
SENT-MOVER	SMD + W2V	0.438	0.505	0.540	0.442	0.514	0.456	0.494	0.484
	SMD + ELMO + PMEANS	0.569	0.558	0.732	0.525	0.581	0.620	0.584	0.595
	SMD + BERT + PMEANS	0.607	0.623	0.770	0.639	0.667	0.641	0.619	0.652
	SMD + BERT + MNLI + PMEANS	0.616	0.643	0.785	0.660	0.664	0.668	0.633	0.667
WORD-MOVER	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + ELMO + PMEANS	0.579	0.588	0.753	0.559	0.617	0.679	0.645	0.631
	WMD-1 + BERT + PMEANS	0.662	0.687	0.823	0.714	0.735	0.734	0.719	0.725
	WMD-1 + BERT + MNLI + PMEANS	0.670	0.708	**0.835**	**0.746**	**0.738**	0.762	**0.744**	**0.743**
	WMD-2 + BERT + MNLI + PMEANS	**0.679**	**0.710**	0.832	0.745	0.736	**0.763**	0.740	**0.743**

Table 1: Absolute Pearson correlations with segment-level human judgments in 7 language pairs on WMT17 dataset.

- Word Mover that use BERT fine tuned on MNLI beats all other metrics.
- It even beats RUSE which is a supervised metric.
- Sentence Mover is not helping. Squeezing the whole sentence in one embedding, and no mapping of words within the sentence as well

Results- Summarization

Setting	Metrics	TAC-2008 Responsiveness r	TAC-2008 Responsiveness ρ	TAC-2008 Pyramid r	TAC-2008 Pyramid ρ	TAC-2009 Responsiveness r	TAC-2009 Responsiveness ρ	TAC-2009 Pyramid r	TAC-2009 Pyramid ρ
BASELINES	S^3_{best} (*)	0.715	0.595	0.754	0.652	0.738	**0.595**	**0.842**	**0.731**
	ROUGE-1	0.703	0.578	0.747	0.632	0.704	0.565	0.808	0.692
	ROUGE-2	0.695	0.572	0.718	0.635	0.727	0.583	0.803	0.694
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
SENT-MOVER	SMD + W2V	0.583	0.469	0.603	0.488	0.577	0.465	0.670	0.560
	SMD + ELMO + PMEANS	0.631	0.472	0.631	0.499	0.663	0.498	0.726	0.568
	SMD + BERT + PMEANS	0.658	0.530	0.664	0.550	0.670	0.518	0.731	0.580
	SMD + BERT + MNLI + PMEANS	0.662	0.525	0.666	0.552	0.667	0.506	0.723	0.563
WORD-MOVER	WMD-1 + W2V	0.669	0.549	0.665	0.588	0.698	0.520	0.740	0.647
	WMD-1 + ELMO + PMEANS	0.707	0.554	0.726	0.601	0.736	0.553	0.813	0.672
	WMD-1 + BERT + PMEANS	0.729	0.595	0.755	0.660	0.742	0.581	0.825	0.690
	WMD-1 + BERT + MNLI + PMEANS	**0.736**	**0.604**	**0.760**	**0.672**	**0.754**	0.594	0.831	0.701
	WMD-2 + BERT + MNLI + PMEANS	0.734	0.601	0.752	0.663	0.753	0.586	0.825	0.694

Table 2: Pearson r and Spearman ρ correlations with summary-level human judgments on TAC 2008 and 2009.

- Responsiveness: Overall content and linguistic quality
- Pyramid Score: how many important semantic content units in reference summary is covered by system summary
- Lexical metric ROUGE perform competitively in this task as compared to other tasks

Results- Image Captioning

Setting	Metric	M1	M2
BASELINES	LEIC(*)	**0.939**	**0.949**
	METEOR	0.606	0.594
	SPICE	0.759	0.750
	BERTSCORE-RECALL	0.809	0.749
SENT-MOVER	SMD + W2V	0.683	0.668
	SMD + ELMO + P	0.709	0.712
	SMD + BERT + P	0.723	0.747
	SMD + BERT + M + P	0.789	0.784
WORD-MOVER	WMD-1 + W2V	0.728	0.764
	WMD-1 + ELMO + P	0.753	0.775
	WMD-1 + BERT + P	0.780	0.790
	WMD-1 + BERT + M + P	**0.813**	**0.810**
	WMD-2 + BERT + M + P	0.812	0.808

Table 4: Pearson correlation with system-level human judgments on MSCOCO dataset. 'M' and 'P' are short names.

- Use the COCO dataset. Systems get M1 and M2 scores for overall quality
- Word Mover beats all other baselines except the supervised metric LEIC, which uses more information by considering both images and text.

Results- Dialogue Response Generation

Setting	Metrics	BAGEL Inf	BAGEL Nat	BAGEL Qual	SFHOTEL Inf	SFHOTEL Nat	SFHOTEL Qual
BASELINES	BLEU-1	0.225	0.141	0.113	0.107	0.175	0.069
	BLEU-2	0.211	0.152	0.115	0.097	0.174	0.071
	METEOR	0.251	0.127	0.116	0.111	0.148	0.082
	BERTSCORE-F1	0.267	0.210	**0.178**	0.163	0.193	0.118
SENT-MOVER	SMD + W2V	0.024	0.074	0.078	0.022	0.025	0.011
	SMD + ELMO + PMEANS	0.251	0.171	0.147	0.130	0.176	0.096
	SMD + BERT + PMEANS	0.290	0.163	0.121	0.192	0.223	0.134
	SMD + BERT + MNLI + PMEANS	0.280	0.149	0.120	0.205	0.239	0.147
WORD-MOVER	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + ELMO + PMEANS	0.261	0.163	0.148	0.147	0.215	0.136
	WMD-1 + BERT + PMEANS	**0.298**	**0.212**	0.163	0.203	0.261	0.182
	WMD-1 + BERT + MNLI + PMEANS	0.285	0.195	0.158	**0.207**	**0.270**	**0.183**
	WMD-2 + BERT + MNLI + PMEANS	0.284	0.194	0.156	0.204	0.270	0.182

Table 3: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.

- Informativeness: how much information does the response give
- Naturalness: how likely is the response from a human
- Quality: How fluent and grammatically correct is the response
- No metric gives a moderately good correlation with human judgement
- Speculate that contextualizers bad at representing named entities.

BERTScore

Approach

- a reference sentence $x = \langle x_1, \ldots, x_k \rangle$
- a candidate sentence $\hat{x} = \langle \hat{x}_1, \ldots, \hat{x}_l \rangle$
- Use contextualized embeddings to represent the token
  - Contrast to word embeddings, contextual embeddings, such as BERT can generate different vector representations for the same word in different sentences depending on the surrounding words
- Compute matching using Cosine Similarity
  - For reference token x_i and candidate token $\hat{x}_j$ the cosine similarity is $\frac{\mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\|\mathbf{x}_i\| \|\hat{\mathbf{x}}_j\|}$
  - They use pre normalized vectors so that denominator need not be computed
- Use greedy matching to maximize matching similarity score

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

- It has been found out that rare words are more indicative for sentence similarity rather than common words
- BERTScore uses inverse document frequency to incorporate importance weighting, calculated over the test corpus

$$\text{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}[w \in x^{(i)}]$$

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \text{idf}(x_i)}$$

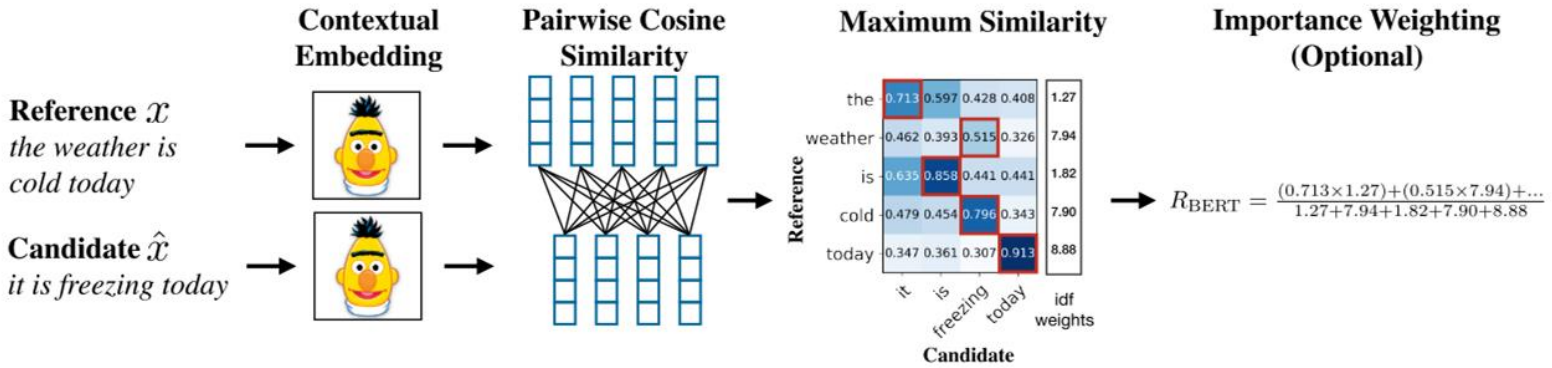


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

- The paper found that the observed values of the metric is coming in a limited range, thus making it less readable.
- To overcome this, rescale the score with respect to its empirical lower bound b, calculated over the common crawl dataset for each embedding model
- Randomly pair 2 sentences from the corpus, compute BERTScore between them.
- Because of the random pairing and the corpus diversity, each pair has very low lexical and semantic overlapping (BLEU score is around 0).
- This gives a minimum value of BERTScore for sentences that are not related
- b is calculated by taking the average over 1M such pairs.

$$\hat{R}_{\text{BERT}} = \frac{R_{\text{BERT}} - b}{1 - b}$$

Results- Machine Translation

- Mainly Evaluated on WMT18 metric evaluation dataset
  - contains predictions of 149 translation systems across 14 language pairs, gold references, and two types of human judgment scores.
  - Segment-level human judgments assign a score to each reference-candidate pair. System-level human judgments associate each system with a single score based on all pairs in the test set.
- Use absolute pearson and kendall rank correlation to evaluate metric quality

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson Correlation

Source: Wikpedia

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables *X* and *Y*, such that all the values of (x_i) and (y_i) are unique (ties are neglected for simplicity). Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be *concordant* if the sort order of (x_i, x_j) and (y_i, y_j) agrees: that is, if either both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$; otherwise they are said to be *discordant*.

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$[3]

Kendall Correlation

Source: Wikpedia

Metric	en↔cs (5/5)	en↔de (16/16)	en↔et (14/14)	en↔fi (9/12)	en↔ru (8/9)	en↔tr (5/8)	en↔zh (14/14)
BLEU	.970/**.995**	.971/**.981**	**.986/.975**	.973/**.962**	.979/**.983**	**.657**/.826	.978/.947
ITER	.975/.915	.990/**.984**	.975/**.981**	**.996/.973**	.937/.975	**.861**/.865	.980/ –
RUSE	.981/ –	.997/ –	**.990**/ –	.991/ –	**.988**/ –	**.853**/ –	**.981**/ –
YiSi-1	.950/**.987**	.992/**.985**	.979/**.979**	.973/.940	**.991/.992**	**.958/.976**	.951/**.963**
P_{BERT}	.980/**.994**	**.998/.988**	**.990/.981**	.995/.957	.982/**.990**	**.791/.935**	.981/.954
R_{BERT}	**.998/.997**	.997/**.990**	.986/**.980**	**.997/.980**	**.995/.989**	.054/.879	**.990/.976**
F_{BERT}	**.990/.997**	**.999/.989**	.990/**.982**	**.998/.972**	**.990**/.990	**.499**/.908	**.988**/.967
F_{BERT} (idf)	.985/**.995**	**.999/.990**	**.992/.981**	.992/**.972**	**.991/.991**	**.826/.941**	**.989/.973**

Table 1: Absolute Pearson correlations with system-level human judgments on WMT18. For each language pair, the left number is the to-English correlation, and the right is the from-English. We bold correlations of metrics not significantly outperformed by any other metric under Williams Test for that language pair and direction. The numbers in parenthesis are the number of systems used for each language pair and direction.

Metric	en↔cs (5k/5k)	en↔de (78k/ 20k)	en↔et (57k/32k)	en↔fi (16k/10k)	en↔ru (10k/22k)	en↔tr (9k/1k)	en↔zh (33k/29k)
BLEU	.233/.389	.415/.620	.285/.414	.154/.355	.228/.330	.145/.261	.178/.311
ITER	.198/.333	.396/.610	.235/.392	.128/.311	.139/.291	-.029/.236	.144/ –
RUSE	.347/ –	.498/ –	.368/ –	.273/ –	.311/ –	.259/ –	.218/ –
YiSi-1	.319/.496	.488/.691	.351/.546	.231/.504	.300/.407	.234/.418	.211/.323
P_{BERT}	.387/.541	.541/.715	.389/.549	.283/.486	.345/.414	.280/.328	.248/.337
R_{BERT}	.388/**.570**	.546/**.728**	.391/**.594**	**.304/.565**	.343/.420	.290/**.411**	.255/**.367**
F_{BERT}	.404/.562	**.550/.728**	**.397**/.586	.296/.546	**.353**/.423	.292/.399	**.264**/.364
F_{BERT} (idf)	**.408**/.553	**.550**/.721	.395/585	.293/.537	.346/**.425**	**.296**/.406	.260/.366

Table 4: Kendall correlations with segment-level human judgments on WMT18. For each language pair, the left number is the to-English correlation, and the right is the from-English. We bold correlations of metrics not significantly outperformed by any other metric under bootstrap sampling for that language pair and direction. The numbers in parenthesis are the number of candidate-reference sentence pairs for each language pair and direction.

Hybrid Systems

- Also experiment with hybrid systems, where they do random sampling for choosing candidate sentence for each reference sentence.
- This allows system level experiments with higher number of systems.
- Human judgements for each hybrid system are created by averaging segment level human judgements for the corresponding sentences in sampled data.

Metric	en↔cs	en↔de	en↔et	en↔fi	en↔ru	en↔tr	en↔zh
BLEU	.956/.993	.969/**.977**	**.981**/.971	.962/.958	.972/.977	.586/.796	.968/.941
ITER	.966/.865	.990/.978	.975/**.982**	.989/.966	.943/.965	.742/.872	.978/ –
RUSE	.974/ –	.996/ –	.988/ –	**.983**/ –	.982/ –	.780/ –	.973/ –
YiSi-1	.942/.985	.991/.983	.976/.976	.964/.938	**.985/.989**	**.881/.942**	.943/.957
P_{BERT}	.965/.989	.995/.983	**.990/.970**	.976/.951	.976/.988	.846/.936	.975/.950
R_{BERT}	**.989/.995**	.997/**.991**	.982/**.979**	.989/**.977**	**.988/.989**	.540/**.872**	**.981/.980**
F_{BERT}	.978/**.993**	.998/.988	.989/.978	.983/.969	.985/.989	.760/.910	**.981**/.969
F_{BERT} (idf)	.982/.995	**.998**/.988	**.988**/.979	**.989**/.969	.983/.987	.453/.877	.980/.963

Pearson Correlation with system level human judgements

Robustness Analysis

- Use adversarial paraphrase classification to test robustness
- Use the Quora Question Pair (QQP) corpus and Paraphrase Adversaries from Word Scrambling (PAWS) dataset
- Positive examples in QQP contain real duplicate questions whereas negative ones are related but different questions
- Sentence pairs in PAWS are generated through word swapping. E.g. ‘Flight from NY to Florida’ and ‘Flight from Florida to NY’ make one pair and a good model should be able to differentiate between the two.

Results

Type	Method	QQP	PAWS_{QQP}
Trained on QQP (supervised)	DecAtt	0.939*	0.263
	DIIN	0.952*	0.324
	BERT	**0.963***	**0.351**
Trained on QQP + PAWS_{QQP} (supervised)	DecAtt	-	0.511
	DIIN	-	0.778
	BERT	-	**0.831**
Metric (Not trained on QQP or PAWS_{QQP})	BLEU	0.707	0.527
	METEOR	0.755	0.532
	ROUGE-L	0.740	0.536
	CHRF++	0.577	0.608
	BEER	0.741	0.564
	EED	0.743	0.611
	CHARACTER	0.698	0.650
	P_{BERT}	0.757	0.687
	R_{BERT}	0.744	0.685
	F_{BERT}	0.761	0.685
	F_{BERT} (idf)	**0.777**	**0.693**

Table 6: Area under ROC curve (AUC) on QQP and PAWS_{QQP} datasets. The scores of trained DecATT (Parikh et al., 2016), DIIN (Gong et al., 2018), and fine-tuned BERT are reported by Zhang et al. (2019). Numbers with * are scores on the held-out test set of QQP. We bold the highest correlations of task-specific and task-agnostic metrics.

Reviews of BERTScore

Pros

- Use contextualized Embeddings [Rohit, Shivangi, Shreya]
- Supports variety of languages (104) [Rohit, Shivangi]
- Token weighting handled through idf [Rohit, Shivangi, Aditya, Vishal]
- Generalized to multiple NL tasks [Rohit, Shivangi]
- High correlation with human metrics [Rohit, Rocktim, Harman, Vishal]
- It is a ‘soft’ measure (wrt BLEU) [Rohit]
  - Not ‘soft’ wrt MoverScore
- Can fine tune the embeddings to specific domains to make the metric adaptive [Shivangi]
- Extensive experimentation [Daman, Rocktim, Shreya]
- highlights the weaknesses of the current metrics using examples [Daman]
- Fully Differentiable [Rocktim, Harman]
- Compute baseline scores used for rescaling [Vishal]

Cons

- More expensive to compute as compared to BLEU [Rohit, Rocktim]
  - Harman Disagrees
- Not optimized for one task [Shivangi]
- Problems of the corpus trained on can make embeddings bad [Shivangi, Shreya]
- Authors do not explain which one of the metrics (P,R,F1,idf) is better and why [Shivangi, Daman, Rocktim]
- Would be good to have experiments on different tasks [Daman]
- Can same embedding model be used for all tasks? [Rocktim]
- Correlation with Human Judgement may not always mean a better score [Harman]
- Curse of dimensionality due to use of cosine similarity [Aditya]
- Not used to evaluate generation [Aditya]
  - What should be the reference sentence here to evaluate on. Another metric perplexity is used.
- Same token can be matched to multiple tokens which might not be good [Vishal]
- Might not work in specialized domains [Vishal]
  - Would have to fine tune the model on a corpus
- Problems with standardization due to hardware dependent outputs of BERT [Vishal]
- Datasets used by the authors have less sequence length [Vishal]

Extensions

- Extend with EMD [Rohit]
  - MoverScore uses this
- Extend it to vision domain [Shivangi, Harman]
- More experimentation [Daman]
- Unified Metric [Daman]
- How would this metric be fooled [Daman]
  - Biases in the embeddings can be a weak point for this metric
- Humans can also do mistakes. Works to analyse these [Harman]
- Need to see how BERTScore can be used with loss function [Harman]

BERTScore vs MoverScore

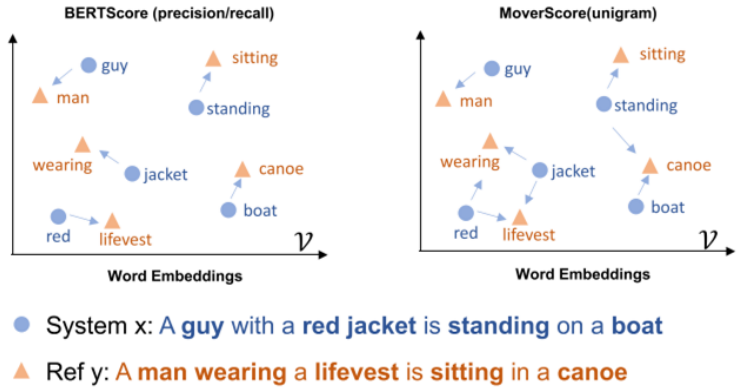
A Comparison

Interesting Times

- MoverScore cites BERTScore
  - “BERTScore (precision/recall) can be represented as a Mover Distance〈C, F〉where C is a transportation cost matrix based on BERT and F is a uniform transportation flow matrix”
- Interestingly, BERTScore also cites MoverScore
  - “They propose various improvements compared to our use of contextualized embeddings. We study these improvements and show that integrating them into BERTSCORE makes it equivalent or better than the mover distance based approach”

What MoverScore talks about BERTScore

- MoverScore can represent BERTScore (precision/recall)
- MoverScore allows to have one to many alignment of words



- It also compares the values of BERTScore in their experiments
- BERTScore has hard alignments. MoverScore has soft alignments

MoverScore Ablation

- Use the representation of the 9th layer only in the experiments and not do aggregation.
- HMD: Hard Mover Distance
- WMD outperforms precision and recall. F1 is competitive.

Metrics	cs-en	de-en	fi-en	lv-en
RUSE	0.624	0.644	0.750	0.697
HMD-F1 + BERT	0.655	0.681	0.821	0.712
HMD-RECALL + BERT	0.651	0.658	0.788	0.681
HMD-PREC + BERT	0.624	0.669	0.817	0.707
WMD-UNIGRAM + BERT	0.651	0.686	**0.823**	0.710
WMD-BIGRAM + BERT	**0.665**	**0.688**	0.821	**0.712**

Table 5: Comparison on hard and soft alignments.

What does BERTScore talk about MoverScore

- Does a more comprehensive ablation study on the different components of MoverScore
  - [PMeans]: The aggregation technique
  - [MNLI]: The dataset it was fine tuned on
  - [IDF-L]: For reference sentences, instead of computing the idf scores on the 560 sentences in the segment-level data ([IDF-S]), compute the idf scores on the 3,005 sentences in the system-level data
  - [SEP]: For candidate sentences, recompute the idf scores on candidate sentences.
  - [RM]: Exclude punctuation marks and all sub word tokens, except the first one, from matching
- Find that PMeans and MNLI fine tuning help the most.

Ablation	Metric	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
Vanilla	WMD1	0.628	0.655	0.795	0.692	0.701	0.715	0.699
	WMD2	0.638	0.661	0.797	0.695	0.700	0.728	0.714
	F_{BERT}	0.659	0.680	0.817	0.702	0.719	0.727	0.717
IDF-S	WMD1	0.636	0.662	0.824	0.709	0.716	0.728	0.713
	WMD2	0.643	0.662	0.821	0.708	0.712	0.732	0.715
	F_{BERT}	0.657	0.681	0.823	0.713	0.725	0.718	0.711
IDF-L	WMD1	0.633	0.659	0.825	0.708	0.716	0.727	0.715
	WMD2	0.641	0.661	0.822	0.708	0.713	0.730	0.716
	F_{BERT}	0.655	0.682	0.823	0.713	0.726	0.718	0.712
IDF-L + SEP	WMD1	0.651	0.660	0.819	0.703	0.714	0.724	0.715
	WMD2	0.659	0.662	0.816	0.702	0.712	0.729	0.715
	F_{BERT}	0.664	0.681	0.818	0.709	0.724	0.716	0.710
IDF-L + SEP + RM	WMD1	0.651	0.686	0.803	0.681	**0.730**	0.730	0.720
	WMD2	0.664	0.687	0.797	0.679	**0.728**	0.735	0.718
	F_{BERT}	0.659	0.695	0.800	0.683	**0.734**	0.722	0.712
IDF-L + SEP + PMEANS	WMD1	0.658	0.663	0.820	0.707	0.717	0.725	0.712
	WMD2	0.667	0.665	0.817	0.707	0.717	0.727	0.712
	F_{BERT}	**0.671**	0.682	0.819	0.708	0.725	0.715	0.704
IDF-L + SEP + MNLI	WMD1	0.659	0.679	0.822	0.732	0.718	0.746	0.725
	WMD2	0.664	0.682	0.819	0.731	0.715	0.748	0.722
	F_{BERT}	0.668	0.701	0.825	**0.737**	0.727	0.744	0.725
IDF-L + SEP + PMEANS + MNLI	WMD1	0.672	0.686	**0.831**	**0.738**	0.725	0.753	**0.737**
	WMD2	**0.677**	0.690	0.828	**0.736**	0.722	**0.755**	0.735
	F_{BERT}	**0.682**	0.707	**0.836**	**0.741**	0.732	0.751	**0.736**
IDF-L + SEP + PMEANS + MNLI + RM	WMD1	0.670	0.708	0.821	0.717	**0.738**	**0.762**	**0.744**
	WMD2	**0.679**	**0.709**	0.814	0.716	**0.736**	**0.762**	0.738
	F_{BERT}	**0.676**	**0.717**	0.824	0.719	**0.740**	**0.757**	**0.738**

Table 9: Ablation Study of MoverScore and BERTScore using Pearson correlations on the WMT17 to-English segment-level data. Correlations that are not outperformed by others for that language pair under Williams Test are bolded. We observe that using WMD does not consistently improve BERTScore.

Not clear that WMD is better than greedy matching or not

Thank you

Any Questions?