291K
Deep Learning for Machine Translation
Processing and Evaluation

Lei Li
UCSB

9/29/2021

Discussion & Homework submission

- Please sign-up yourself at https://piazza.com/class/ksousnwx3cl1ux (we use free version of piazza, you may see piazza donation banner)
- 5 points for active discussion and sharing experience.
- Turn-in your home at https://www.gradescope.com/courses/319418
  - HW1 has two separate submissions, one for pdf file, the other for problem 3&4 coding (zip file)
  - Due date listed on Course website and also on Gradescope
- Sign-up for HW4: language presentation
  - https://tinyurl.com/4m8yjkuv (avoid grouping with same person in project)
  - Prefer low-resource languages.

Outline

- Corpus resource
  - Text Corpus: Parallel, Monolingual, Document-level
- Vocabulary building & Tokenization
- Evaluation
  - Automatic metric
  - Human evaluation

Commonly-used (Text) Machine Translation data

- (Rich-resource) WMT 14 En-De: http://statmt.org/wmt14/translation-task.html#Download
  - tool to download: https://github.com/bytedance/neurst/blob/master/examples/translation/download_wmt14en2de.py
- (Low-resource) WMT 16 En-Ro: https://www.statmt.org/wmt16/translation-task.html#download

Dataset	WMT 14 En-De	WMT16 En-Ro
Parallel	4.5m	0.62m
Non-parallel	5m	1m
Dev	newstest2013	newstest2015
Test	newstest2014	newstest2016

Vocabulary

- To model P(y|x)
- Consider a ten-word sentence, chosen from common English dictionary about 5k words
  - 5000^{10} possible sentences
  - need a table of $5000^{10} \cdot 5000^{10}$ entries, infeasible
- source and target sentences need to break into smaller units.
- Multiple ways to segment
- Language specific considerations

Tokenization

- Break sentences into tokens, basic elements of processing
- Word-level Tokenization
  - Break by space and punctuation.
  - English, French, German, Spanish

The	most	eager	is	Oregon	which	is	enlisting	5,000	drivers	in	the	country	’s	biggest	experiment	.

  - Special treatment: numbers replaced by special token [number]
  - How large is the Vocabulary? Cut-off by frequency, the rest replaced by [UNK]

Pros and Cons of Word-level Tokenization

- Easy to implement
- Cons:
  - Out-of-vocabulary (OOV) or unknown tokens, e.g. Covid
  - Tradeoff between parameters size and unknown chances.
    - Smaller vocab => fewer parameters to learn, easier to generate (deciding one word from smaller dictionary), more OOV
    - Larger vocab => more parameters to learn, harder to generate, less OOV
  - Hard for certain languages with continuous script: Japanese, Chinese, Korean, Khmer, etc. Need separate word segmentation tool (can be neural networks)

最热切的是俄勒冈州，该州正在招募 5,000 名司机参与该国最大的试验。

Character-level Tokenization

T	h	e	m	o	s	t	e	a	g	e	r	i	s	O	r	e	g	…

- Each letter and punctuation is a token
- Pros:
  - Very small vocabulary (except for some languages, e.g. Chinese)
  - No Out-of-Vocabulary token
- Cons:
  - A sentence can be longer sequence
  - Tokens do not representing semantic meaning

Subword-level Tokenization

The most eager is Oregon which is en listing 5,000 driver s in the country ’s big g est experiment .

- Goal:
  - moderate size vocabulary
  - no OOV
- Idea:
  - represent rare words (OOV) by sequence of subwords
- Byte Pair Encoding (BPE)
  - not necessarily semantic meaningful
  - Originally for data compression

Philip Gage. A New Algorithm for Data Compression, 1994

Byte Pair Encoding

- Use smallest sequence of strings to represent original string. Group frequent pair of bytes together.
- Put all characters into symbol table
- For each loop, until table reach size limit
  - count frequencies of symbol pair
  - replace most frequent pair with a new symbol, add to symbol table

Byte Pair Encoding (BPE) for Text Tokenization

1. Initialize vocabulary with all characters as tokens (also add end-of-word symbol) and frequencies
2. Loop until vocabulary size reaches capacity
   1. Count successive pairs of tokens in corpus
   2. Rank and select the top frequent pair
   3. Combine the pair to form a new token, add to vocabulary
3. Output final vocabulary and tokenized corpus

Rico Sennrich et al. Neural Machine Translation of Rare Words with Subword Units. 2016

Example

l, o, w, e, r, n, s, t, i, d, </w>	‘l o w</w>’ : 5 ‘l o w e r</w>’ : 2 ’n e w e s t</w>’ : 6 ‘w i d e s t</w>’ : 3
l, o, w, e, r, n, s, t, i, d, </w>, es	‘l o w</w>’ : 5 ‘l o w e r</w>’ : 2 ’n e w es t</w>’ : 6 ‘w i d es t</w>’ : 3
l, o, w, e, r, n, s, t, i, d, </w>, es, est	‘l o w</w>’ : 5 ‘l o w e r</w>’ : 2 ’n e w est</w>’ : 6 ‘w i d est</w>’ : 3
l, o, w, e, r, n, s, t, i, d, </w>, es, est, est</w>	‘l o w</w>’ : 5 ‘l o w e r</w>’ : 2 ’n e w est</w>’ : 6 ‘w i d est</w>’ : 3
l, o, w, e, r, n, s, t, i, d, </w>, es, est, est</w>, lo,	‘lo w</w>’ : 5 ‘lo w e r</w>’ : 2 ’n e w est</w>’ : 6 ‘w i d est</w>’ : 3
l, o, w, e, r, n, s, t, i, d, </w>, es, est, est</w>, lo, low	‘low</w>’ : 5 ‘low e r</w>’ : 2 ’n e w est</w>’ : 6 ‘w i d est</w>’ : 3

More Subword Tokenization

- Wordpiece:
  - like BPE
  - but instead of merge with most frequent pairs, merge a and b, if p(b|a) will be maximized
- SentencePiece:
  - Uniform way to treat space, punctuation
  - Use the raw sentence, replacing space ‘ ’ with _ (U+2581)
  - Then split character and do BPE

Kudo and Richardson, SentencePiece, 2018

Many possible translation, which is better?

SpaceX周三晚间进行了一次发射任务，将四名毫无航天经验的业余人士送入太空轨道。

SpaceX launched a mission Wednesday night to put four amateurs with no space experience into orbit.

SpaceX conducted a launch mission on Wednesday night, sending four amateurs with no aerospace experience into space orbit.

SpaceX conducted a launch mission Wednesday night that sent four amateurs with no spaceflight experience into orbit.

SpaceX carried out a launch mission on Wednesday night to put four amateurs without Aerospace experience into orbit.

Assessing the Quality of Translation

- Criteria for evaluation metric
  - Consistent across different evaluation, so that translation quality is comparable
  - Differentiable: tell high quality translation from low quality ones
  - Low cost: requires low effort of human (e.g. amateur can perform) or computation
-

Aspects of Translation Quality

- Intuition
  - Scoring of translations is (implicitly) based on an identification of errors and other imperfections.
- Adequacy/Faithfulness
  - Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
- Expressiveness
- Elegance
- Due to Yan Fu (1854-1921)

Direct Assessment of Translation Quality

- Source-based
  - Human annotators are given source, without reference.
  - avoid bias
  - can also be used to evaluate human translation performance
- Reference-based
  - Human annotators are given reference, without source.
  - Can be done by monolingual speaker in target language
  - Less effort
- Source-Reference

Direct Assessment of Translation Quality

- Grading scheme
  - 1-4, 1-5, 1-6
  - 0-100 scale (used in WMT 2020)
- Does it require professional translator or amateur(college students in Foreign language)

4	Correct translation and fluent language
3	Mostly understandable, with 1 or 2 errors
2	some meaningful, but more errors
1	incorrect or major errors

WMT 2020 Evaluation

- 2887 Turkers recruited on Amazon Mechanical Turk.
- 2233 are removed, not passing the quality control
- 654 Turkers are adopted
- 166,868 assessment scores (of 654k)
- For 10 to-English pairs (Chinese, Czech, German, Russian, etc.)
- Turkers are provided source and machine translated output
- Quality Control (next)

Barrault et al. Findings of the 2020 Conference on Machine Translation (WMT20), 2020

Quality Control

- How to ensure that crowd raters produce high quality assessment?
- 100 translation assessment: 40 are regular
- Repeat pairs (10): expecting similar judgement
- Bad Reference Pairs (10):
  - damaged MT outputs by randomly replacing n-gram phrases from the same test set.
  - expects low scores
- Good Reference Pairs (10)
  - Use golden reference
  - expects high scores
- Excluding Bad (10) and Good (10) in calculating final score.

Filtering Low-quality Annotators

- How to tell if an annotator consistently scores bad references pairs lower?
- Hypothesis testing (significance test)
  - Annotator scores MT pair with X
  - Annotator scores Bad Reference Pair Y
  - Y < X
  - Is the annotator reliable in assessment? (Is the difference statistically significant?)
- Remove annotators whose scores for normal MT not different from bad reference pairs!

Hypothesis Testing

- Null hypothesis
  - assumption that there is no real difference
- P-Levels
  - probability that the null hypothesis is true
  - p-level $p < 0.01$ = more than 99% chance that difference is real
  - typically used: p-level 0.05 or 0.01
- Confidence Intervals
  - given that the measured score is x
  - what is the true score (on a infinite size test set)?
  - interval $[x - d, x + d]$ contains true score with, e.g., 95% probability

Is the score of system A better than B?

- n pairs of (e.g. MT output, degraded bad translation)
- Scores from human annotators for each (x_i, y_i)
- Null Hypothesis:

  ui=xi - yi is close to 0
- Test statistic:

  $t = \dfrac{\bar{u}}{s/\sqrt{n}}$, where mean difference $\bar{u} = \dfrac{u_i}{n} = \dfrac{x_i - y_i}{n}$,

  standard deviation: $s = \sqrt{\dfrac{1}{n-1}(u_i - \bar{u})^2}$
- e.g. WMT20, n is 10 (for one 100-item batch)
- Compare with t-distribution table: T=1.645 for p-value 0.05

Alternative Annotator Agreement

- For discrete scores (e.g. 1-4)
- Kappa coefficient

$$\kappa = \frac{p(A) - p_r}{1 - p_r}$$

- p(A): percentage of agreed assessments
- p_r: percentage of agreement if random guess (=1/K if there K discrete labels)
- e.g. P(A) = 0.4, P_r=0.25, k=0.2

Ranking and Annotator Difference

- In WMT20, scores of a same annotators are normalized by according to mean and standard deviation
- The overall score is an average of standardized scores.
- Ranking based on overall-score (avg z)

Example Results from WMT 20

Chinese→English

Ave.	Ave. z	System
77.5	0.102	VolcTrans
77.6	0.089	DiDi-NLP
77.4	0.077	WeChat-AI
76.7	0.063	Tencent-Translation
77.8	0.060	Online-B
78.0	0.051	DeepMind
77.5	0.051	OPPO
76.5	0.028	THUNLP
76.0	0.016	SJTU-NICT
72.4	0.000	Huawei-TSC
76.1	−0.017	Online-A
74.8	−0.029	HUMAN
71.7	−0.071	Online-G
74.7	−0.078	dong-nmt
72.2	−0.106	zlabs-nlp
72.6	−0.135	Online-Z
67.3	−0.333	WMTBiomedBaseline

English→Chinese

Ave.	Ave. z	System
80.6	0.568	HUMAN-B
82.5	0.529	HUMAN-A
80.0	0.447	OPPO
79.0	0.420	Tencent-Translation
77.3	0.415	Huawei-TSC
77.4	0.404	NiuTrans
77.7	0.387	SJTU-NICT
76.6	0.373	VolcTrans
73.7	0.282	Online-B
73.0	0.241	Online-A
69.5	0.136	dong-nmt
68.5	0.135	Online-Z
70.1	0.122	Online-G
68.7	0.082	zlabs-nlp

Example Results from WMT 20

Japanese→English

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

English→Japanese

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z
53.9	−0.132	SJTU-NICT
52.8	−0.164	Online-G

Example Results from WMT 20

German→English

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B
78.5	0.131	UEDIN
78.8	0.085	Online-Z
74.2	−0.079	WMTBiomedBaseline
71.1	−0.106	zlabs-nlp
20.5	−1.618	yolo

English→German

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	−0.120	Online-G
68.1	−0.278	zlabs-nlp
65.5	−0.338	WMTBiomedBaseline

Example Results from WMT 20

German→ French

Ave.	Ave. z	System
90.4	0.279	OPPO
90.2	0.266	VolcTrans
89.7	0.262	IIE
89.2	0.243	HUMAN
89.1	0.226	Online-B
89.1	0.223	Online-A
88.5	0.208	Online-G

French → German

Ave.	Ave. z	System
89.8	0.334	VolcTrans
89.7	0.333	OPPO
89.1	0.319	IIE
89.0	0.295	Online-B
87.4	0.247	HUMAN
87.3	0.240	Online-A
87.1	0.221	SJTU-NICT
86.8	0.195	Online-G
85.6	0.155	Online-Z

Expert Rating - MQM

- Multidimensional Quality Metrics
- Rate with error category and severity level
- Error Category: Accuracy, Fluency, Terminology, Style, and Locale

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Freitag et al, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, 2021

MQM Error Category

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize the 5 most severe errors.

Automatic Metric

- The need of automatic metric:
  - Human evaluation is expensive
  - Need fast turnaround for model development
- Easy for text classification, just comparing one label
- Hard for variable-length sequence
  - multiple yet correct translation
- Widely adopted metric: BLEU
  - BiLingual Evaluation Understudy

Word Error Rate

- Minimum number of editing steps to transform output to reference
  - match: words match, no cost
  - substitution: replace one word with another
  - insertion: add word
  - deletion: drop word
- Levenshtein distance

$$\frac{\#substitution + \#insertion + \#deletion}{reference.length}$$

BLEU

- Measuring the precision of n-grams
  - Precision of n-gram: percentage of tokens in output sentences

$$\_\, p_n = \frac{num\,.\,of\,.\,correct\,.\,token\,.\,ngram}{total\,.\,output\,.\,ngram}$$

- Penalize for brevity
  - if output is too short
  - $bp = min(1, e^{1-r/c})$
- BLEU=$bp \cdot (\prod p_i)^{\frac{1}{4}}$
- Notice BLEU is computed over the whole corpus, not on one sentence

Example

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

System B: A rocket sent SpaceX into orbit Wednesday.

Example

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

	Precision
Unigram	9/11
Bigram	4/10
Trigram	2/9
Four-gram	1/8

$$bp = e^{1-12/11} = 0.91$$

$$BLEU = 0.91 * (9/11 * 4/10 * 2/9 * 1/8)^{1/4}$$

$$= 28.1\%$$

Exercise: Calculate BLEU

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System B: A rocket sent SpaceX into orbit Wednesday.

Multi-BLEU

- To account for variability if one source has multiple references.
- Precision
  – n-grams can match in any of the references

  – $p_n = \dfrac{num.of.correct.token.ngram}{total.output.ngram}$
- Brevity Penalty
  – $bp = min(1, e^{1-r/c})$
  – closest reference length used
- BLEU=$bp \cdot (\prod p_i)^{\frac{1}{4}}$
- Notice BLEU is computed over the whole corpus, not on one sentence

Pitfall in Calculating BLEU

- Be careful! Tokenization and normalization make diff!

Ref: A SpaceX rocket was launched into a space orbit Wednesday evening.

System A: SpaceX launched a mission Wednesday evening into a space orbit.

- What is the BLEU for Char-level Tokenization:

Ref: A S p a c e X r o c k e t w a s l a u n c h e d i n t o a s p a c e o r b i t W e d n e s d a y e v e n i n g .

System A: S p a c e X l a u n c h e d a m i s s i o n W e d n e s d a y e v e n i n g i n t o a s p a c e o r b i t .

BLEU scores can differ much!

Data from WMT17 for the same system output using different BLEU configuration.

config	en-cs	en-de	en-fi	en-lv	en-ru	en-tr	cs-en	de-en	fi-en	lv-en	ru-en	tr-en
	English→ ⋆						⋆ →English					
basic	20.7	25.8	22.2	16.9	33.3	18.5	26.8	31.2	26.6	21.1	36.4	24.4
split	20.7	26.1	22.6	17.0	33.3	18.7	26.9	31.7	26.9	21.3	36.7	24.7
unk	20.9	26.5	25.4	18.7	33.8	20.6	26.9	31.4	27.6	22.7	37.5	25.2
metric	20.1	26.6	22.0	17.9	32.0	19.9	27.4	33.0	27.6	22.0	36.9	25.6
range	0.6	0.8	0.6	1.0	1.3	1.4	0.6	1.8	1.0	0.9	0.5	1.2
basic_{lc}	21.2	26.3	22.5	17.4	33.3	18.9	27.7	32.5	27.5	22.0	37.3	25.2
split_{lc}	21.3	26.6	22.9	17.5	33.4	19.1	27.8	32.9	27.8	22.2	37.5	25.4
unk_{lc}	21.4	27.0	25.6	19.1	33.8	21.0	27.8	32.6	28.3	23.6	38.3	25.9
metric_{lc}	20.6	27.2	22.4	18.5	32.8	20.4	28.4	34.2	28.5	23.0	37.8	26.4
$\textit{range}_{lc}$	0.6	0.9	0.5	1.1	0.6	1.5	0.7	1.7	1.0	1.0	0.5	1.2

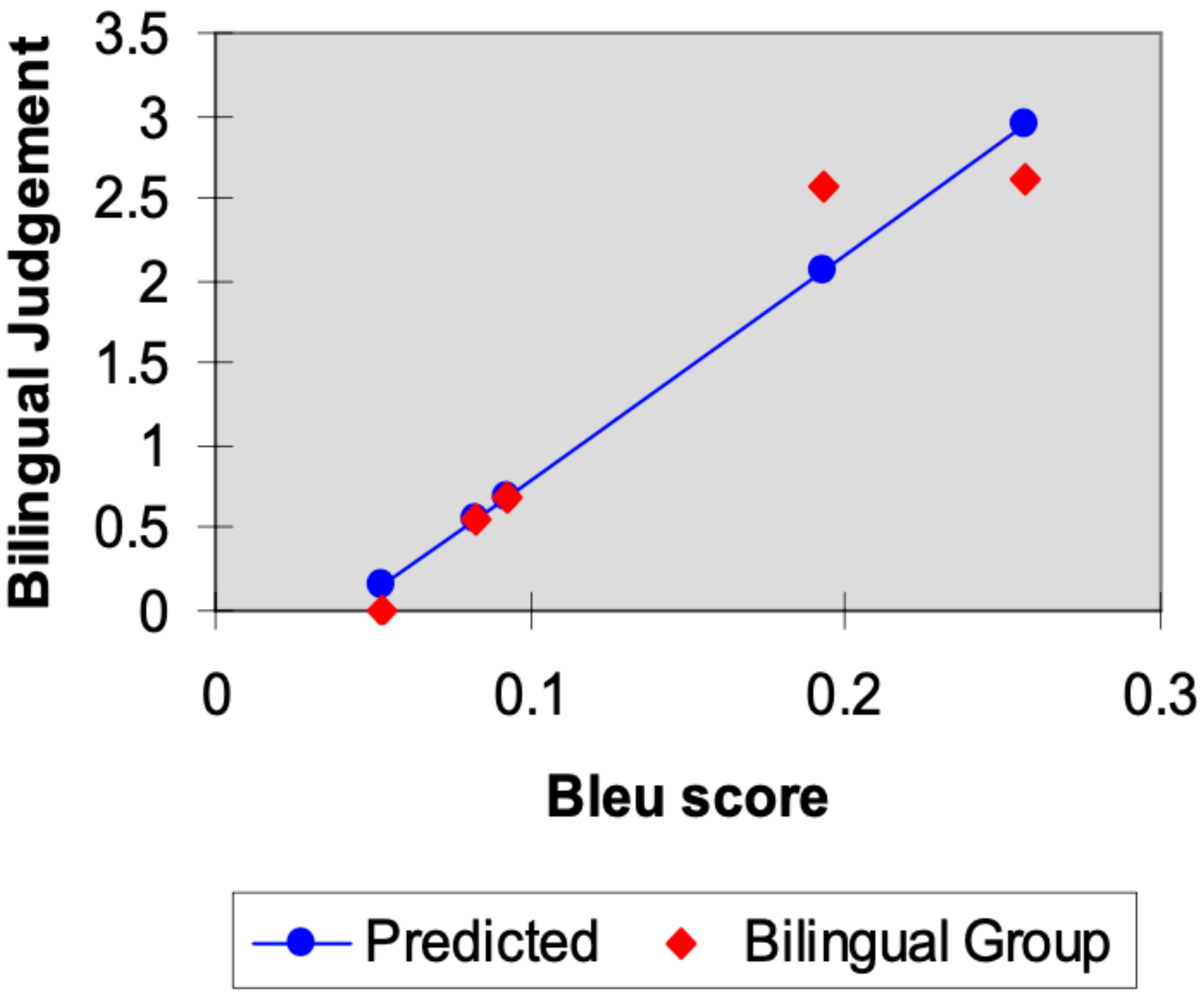
Matt Post. A Call for Clarity in Reporting BLEU Scores, 2018

Guideline of Using BLEU

- Always use sacreBLEU to report
  - also known as detokenized BLEU
  - use metric's original tokenization, no processing on the reference data!!!
    - because different way to tokenize, whether to split compound words (e.g. long-term ==> long - term), cased or uncased can all affect BLEU

Is BLEU correlated with Human Evaluation?

Figure 6: BLEU predicts Bilingual Judgments



Papenani et al, BLEU: a Method for Automatic Evaluation of Machine Translation. 2002

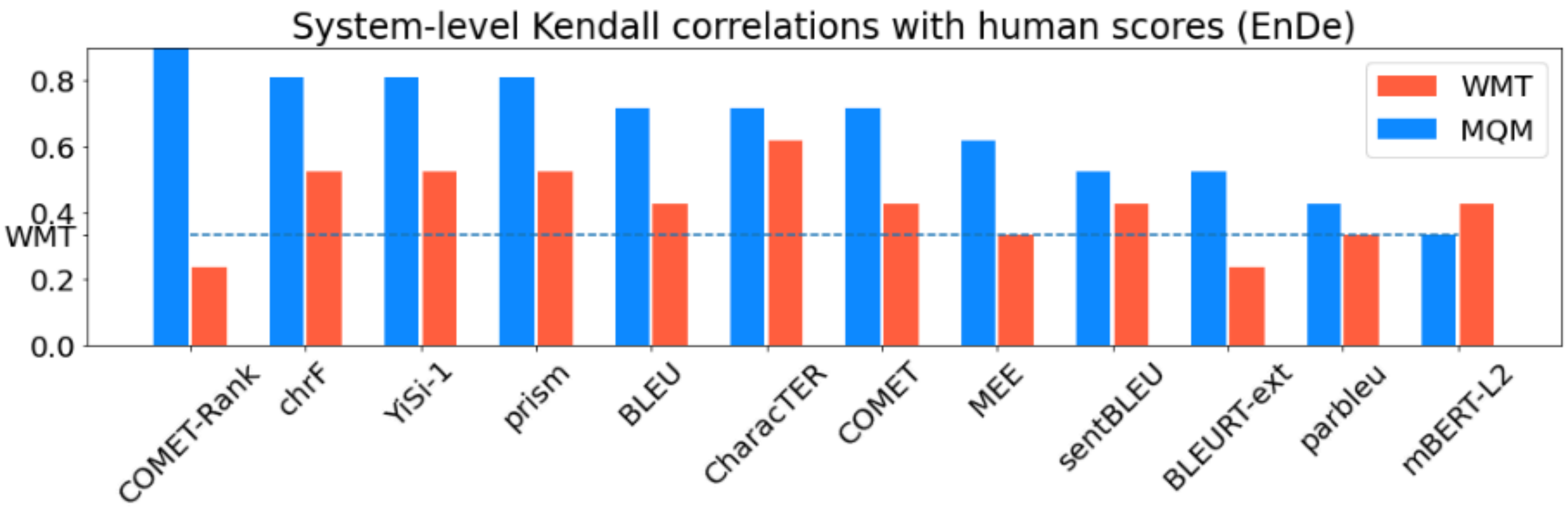
Other Metric

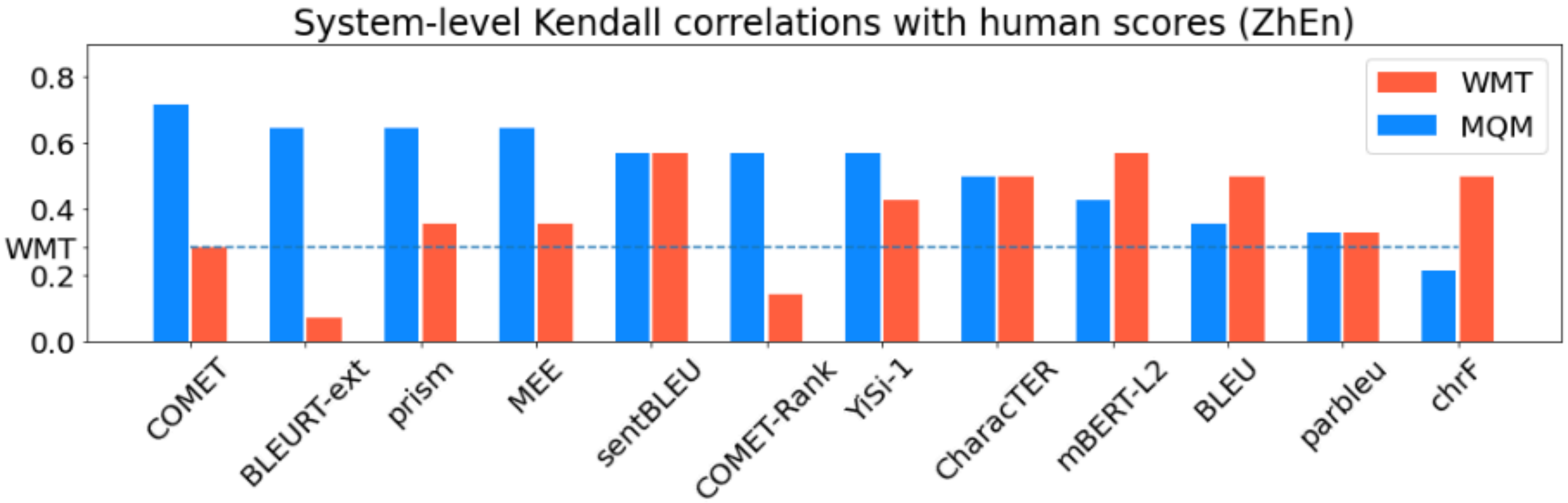
- METEOR: penalty adjusted harmonic average on precision and recall
- penalty $Pen = \gamma(\frac{ch}{m})^{\beta}$, ch is number of matched chunks, m is matched tokens,
- Precision and Recall as before
- $Score = (1 - Pen)\frac{P \cdot R}{\alpha \cdot P + (1-\alpha)R}$
- e.g. $\gamma = 0.5, \beta = 3, \alpha = 0.9$

Learned Metrics

- Use a machine learning model to measure the quality of translation
- e.g. COMET, BERT-score
- prism: using a learned paraphrase model
- Will revisit after next few lectures

Automatic Learned Metric can be good!





Freitag et al, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, 2021

Reference

- Freitag et al, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, 2021
- Philip Koehn. Statistical Significance Tests for Machine Translation Evaluation, 2004
- Papenani et al, BLEU: a Method for Automatic Evaluation of Machine Translation. 2002
- Matt Post. A Call for Clarity in Reporting BLEU Scores, 2018
- Rico Sennrich et al. Neural Machine Translation of Rare Words with Subword Units. 2016
- Barrault et al. Findings of the 2020 Conference on Machine Translation (WMT20), 2020