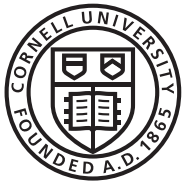
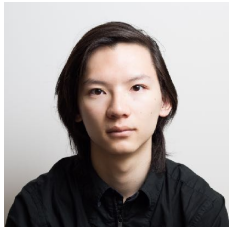


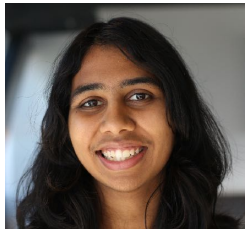
Cornell University

ASAPP

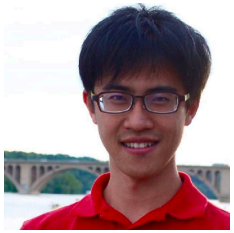
BERTScore: Evaluating Text Generation with BERT



Tianyi Zhang



Varsha Kishore



Felix Wu

Kilian Q. Weinberger

Yoav Artzi

ich liebe es

translate

I am like

I like

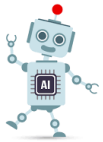
I like it

I love it

I am loving it

ich liebe es

translate



Candidate

I like it

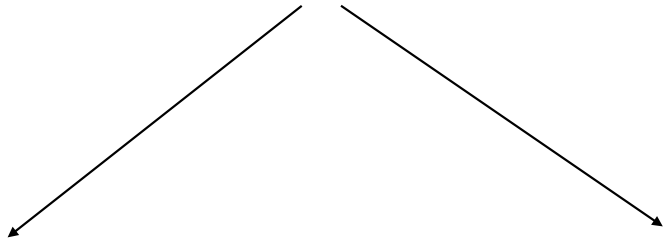
Reference

I love it

0.88/1.00

Metric

Text Generation Evaluation Metrics



N-gram matching approaches

- BLEU (Papineni et al., 2002)
- METEOR (Banerjee & Lavie, 2005)
- ROUGE (Lin, 2004)
- chrF (Popovic, 2015)

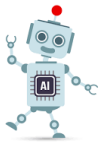
Embedding-based metrics

- Meant 2.0 (Lo, 2017)
- YiSi -1 (Lo et al., 2018)
- BERTScore

BLEU N-gram Matching

Reference

The weather is cold today



Candidate 1

The weather is sunny today

Candidate 2

It is freezing today

BLEU cannot identify synonyms

BLEU gives higher score to candidate 1

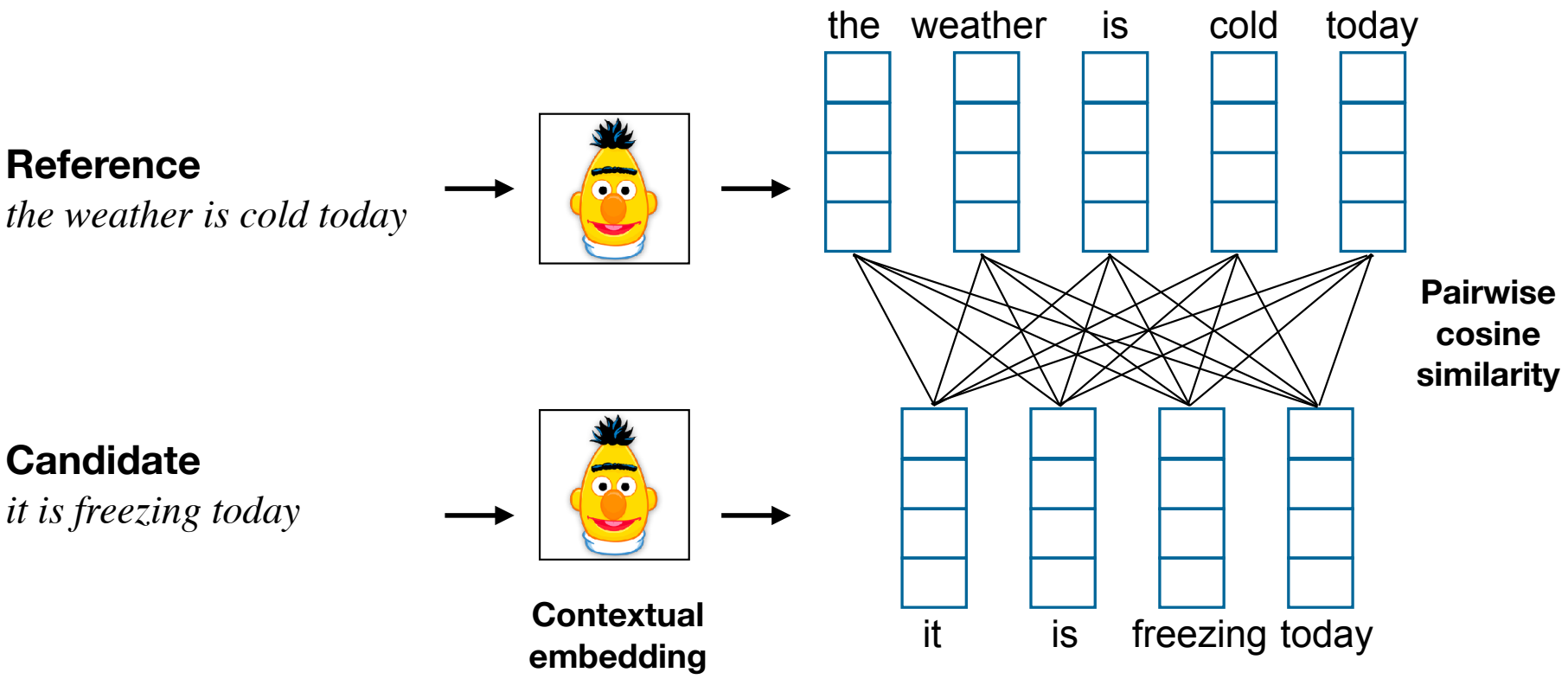
BERTScore: an evaluation metric that uses BERT embeddings

BERT

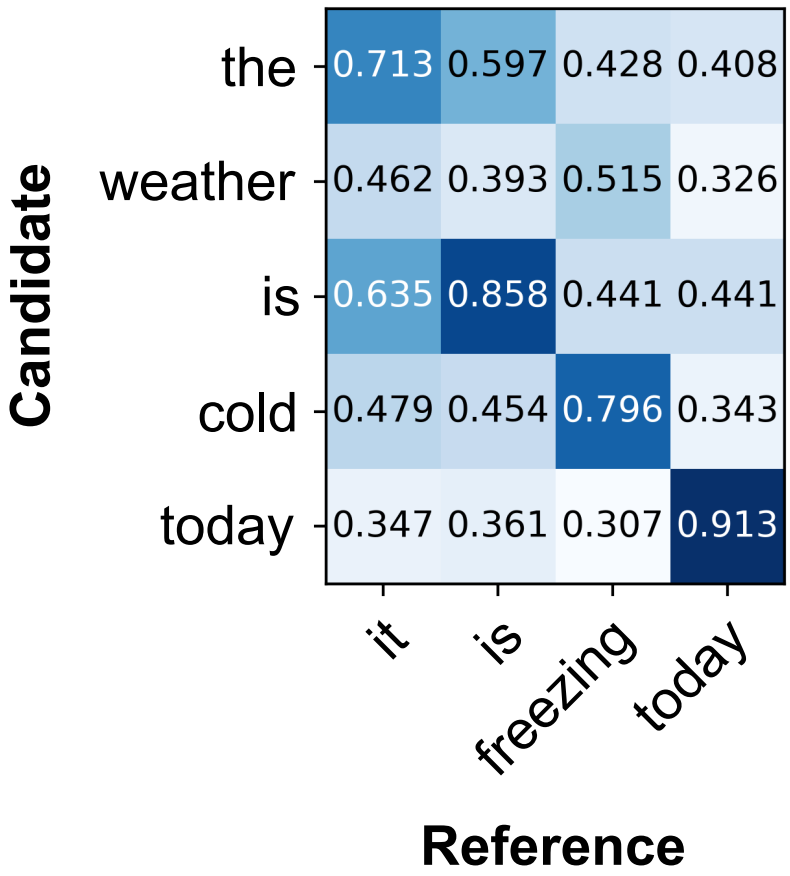
Transformer model
pre-trained on
masked language modeling
and next sentence prediction

Generates word token
embeddings that reflect
their context

BERTScore

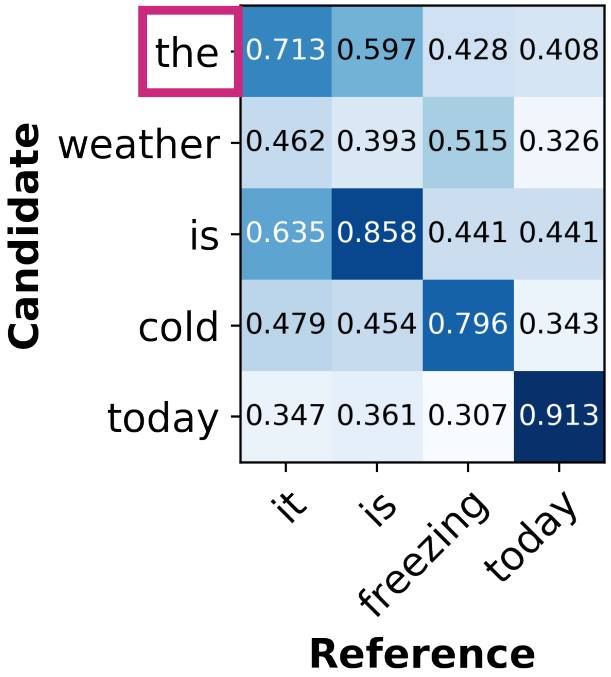

Reference
the weather is cold today
Candidate
it is freezing today
Contextual embedding
the
weather
is
cold
today
it
is
freezing
today
Pairwise cosine similarity

Greedy Matching



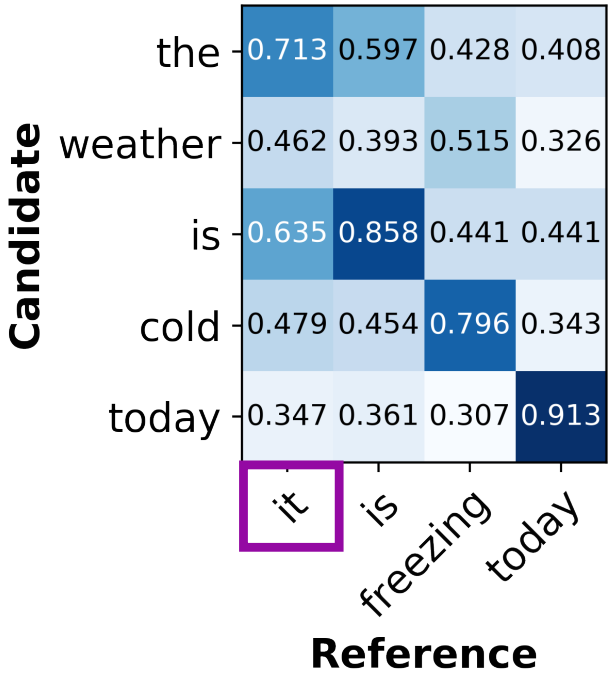
Greedy Matching

Precision



Match words in candidate to reference

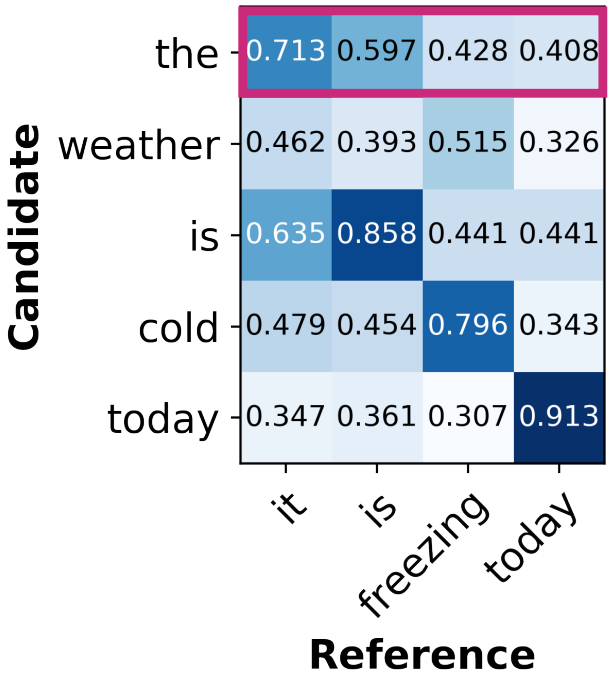
Recall



Match words in reference to candidate

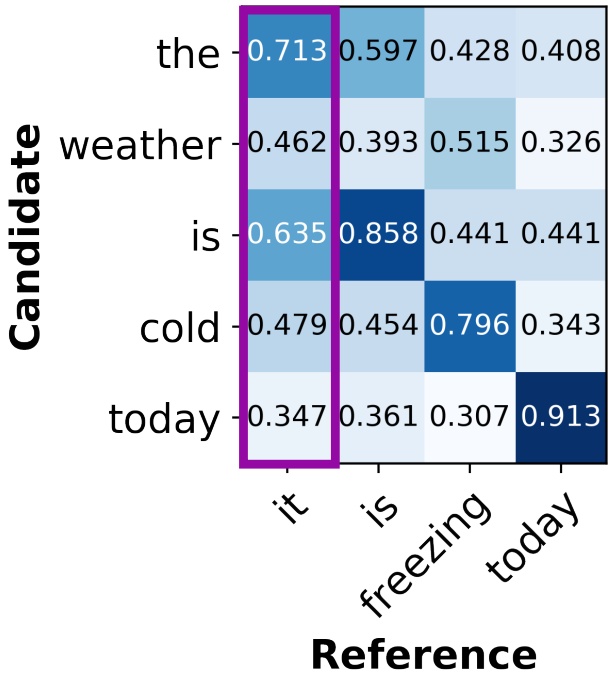
Greedy Matching

Precision



Match words in candidate to reference

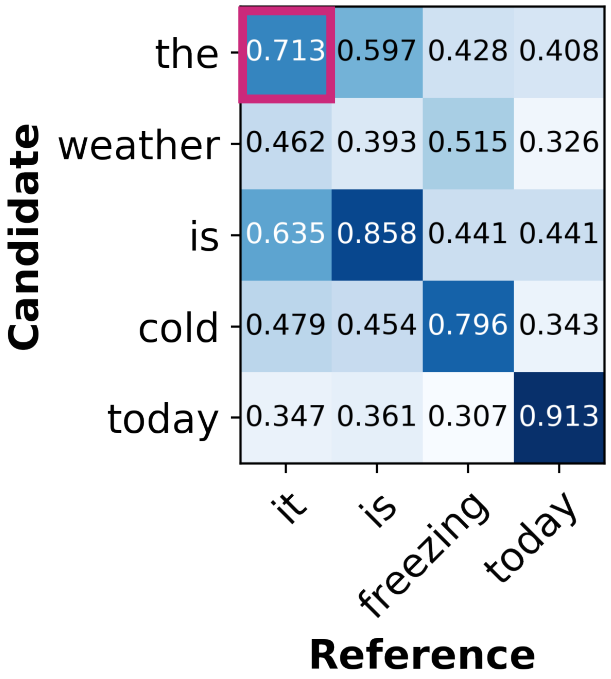
Recall



Match words in reference to candidate

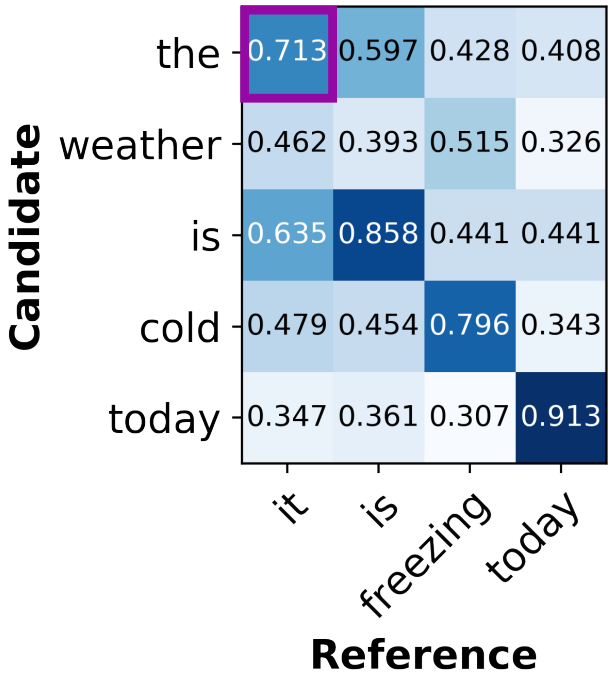
Greedy Matching

Precision



Match words in candidate to reference

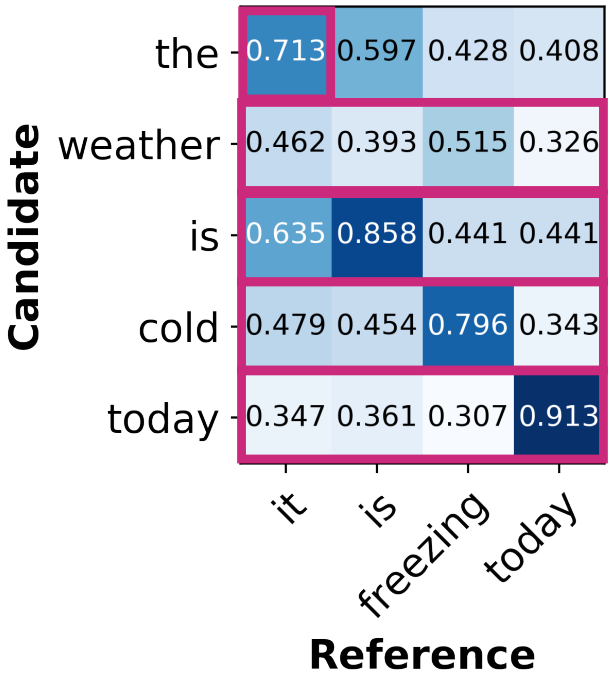
Recall



Match words in reference to candidate

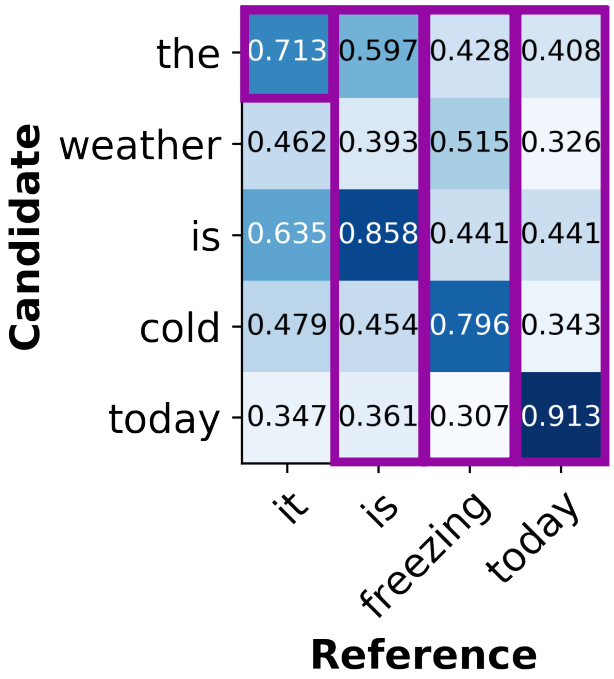
Greedy Matching

Precision



Match words in candidate to reference

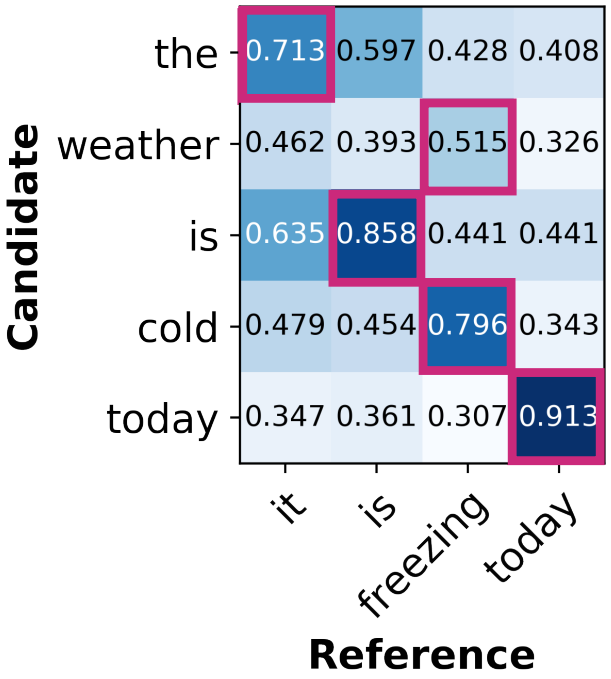
Recall



Match words in reference to candidate

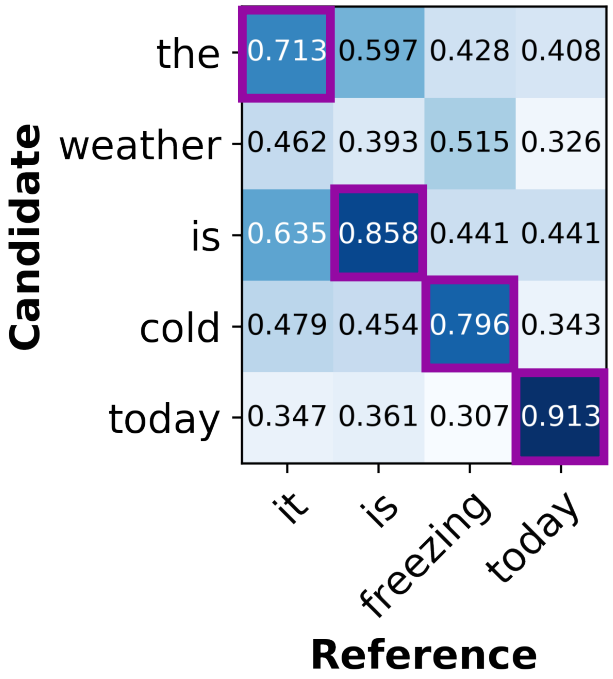
Greedy Matching

Precision



Match words in candidate to reference

Recall

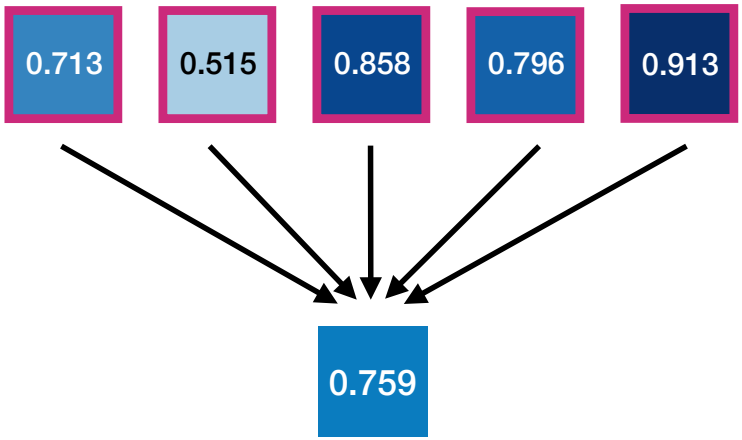


Match words in reference to candidate

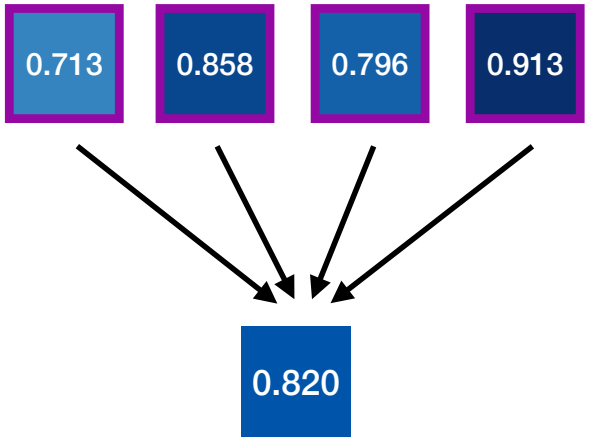
Greedy Matching - Aggregate

Greedy Matching - Aggregate

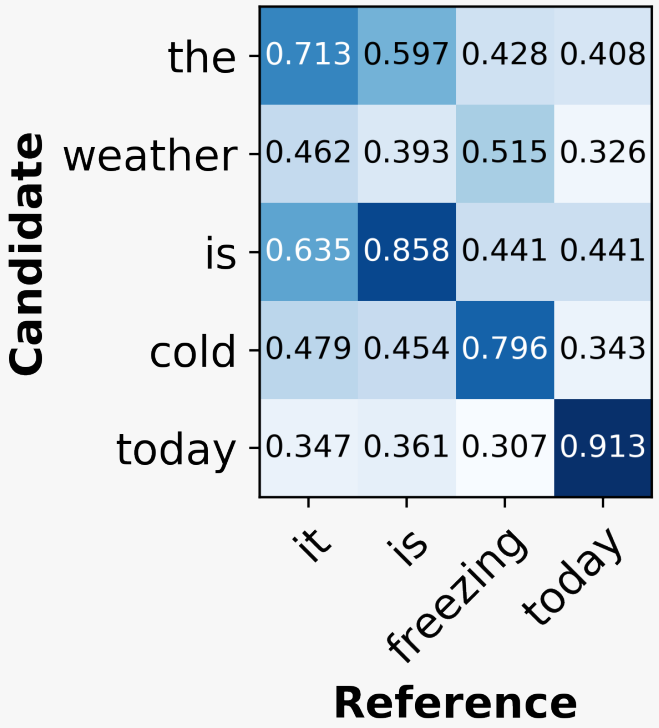
Precision

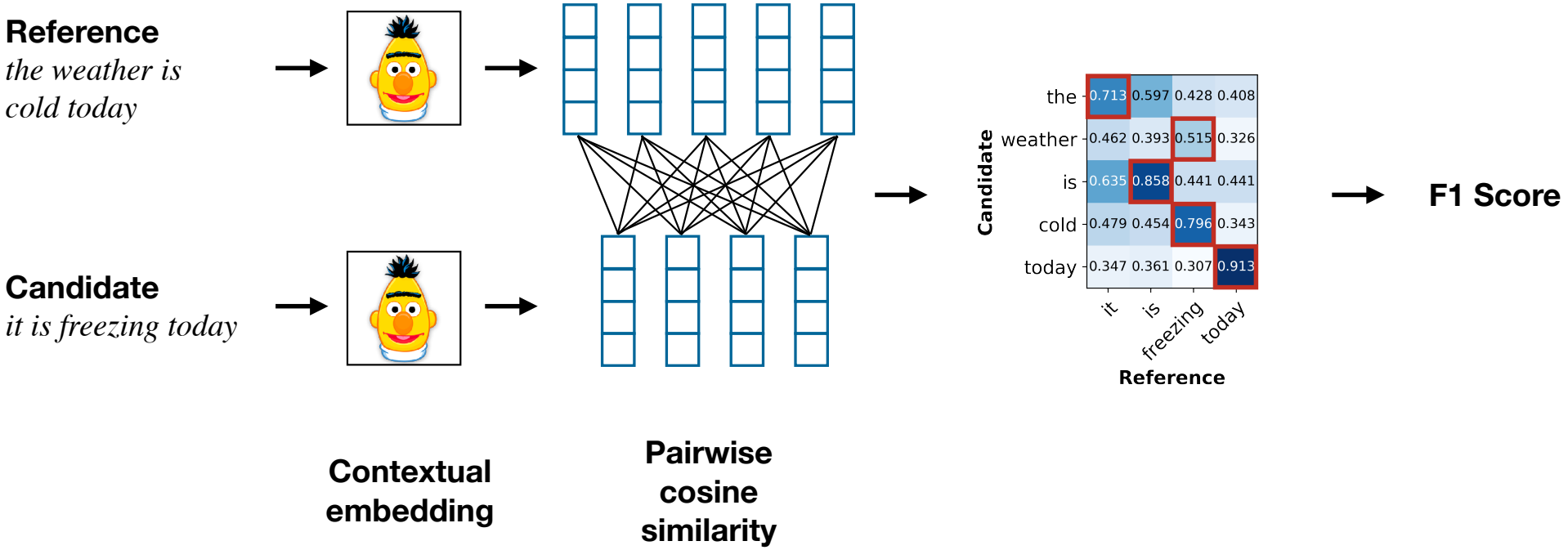


Recall



$$\mathrm{F1} = 2\,\frac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$



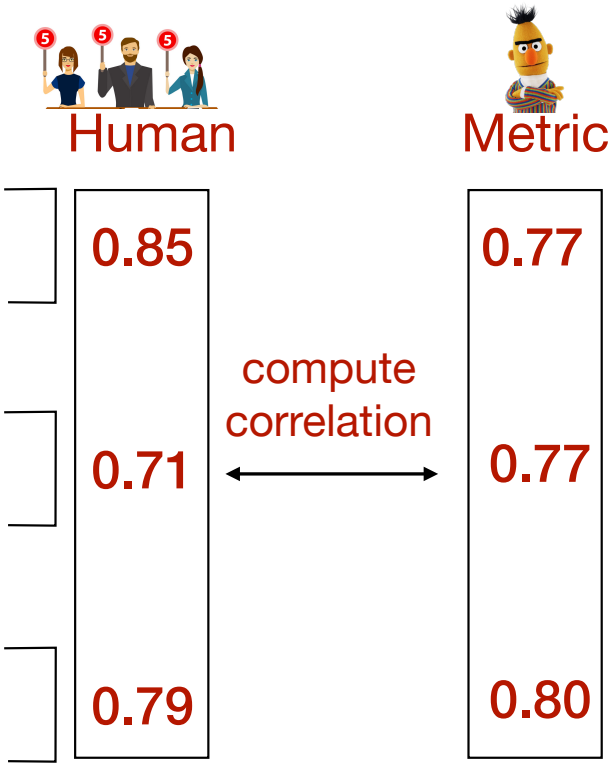

Reference
the weather is cold today
Candidate
it is freezing today
Contextual embedding
Pairwise cosine similarity
Candidate
Reference
the
weather
is
cold
today
it
is
freezing
today
0.713 0.597 0.428 0.408
0.462 0.393 0.515 0.326
0.635 0.858 0.441 0.441
0.479 0.454 0.796 0.343
0.347 0.361 0.307 0.913
F1 Score

Evaluation: WMT Translation Benchmark

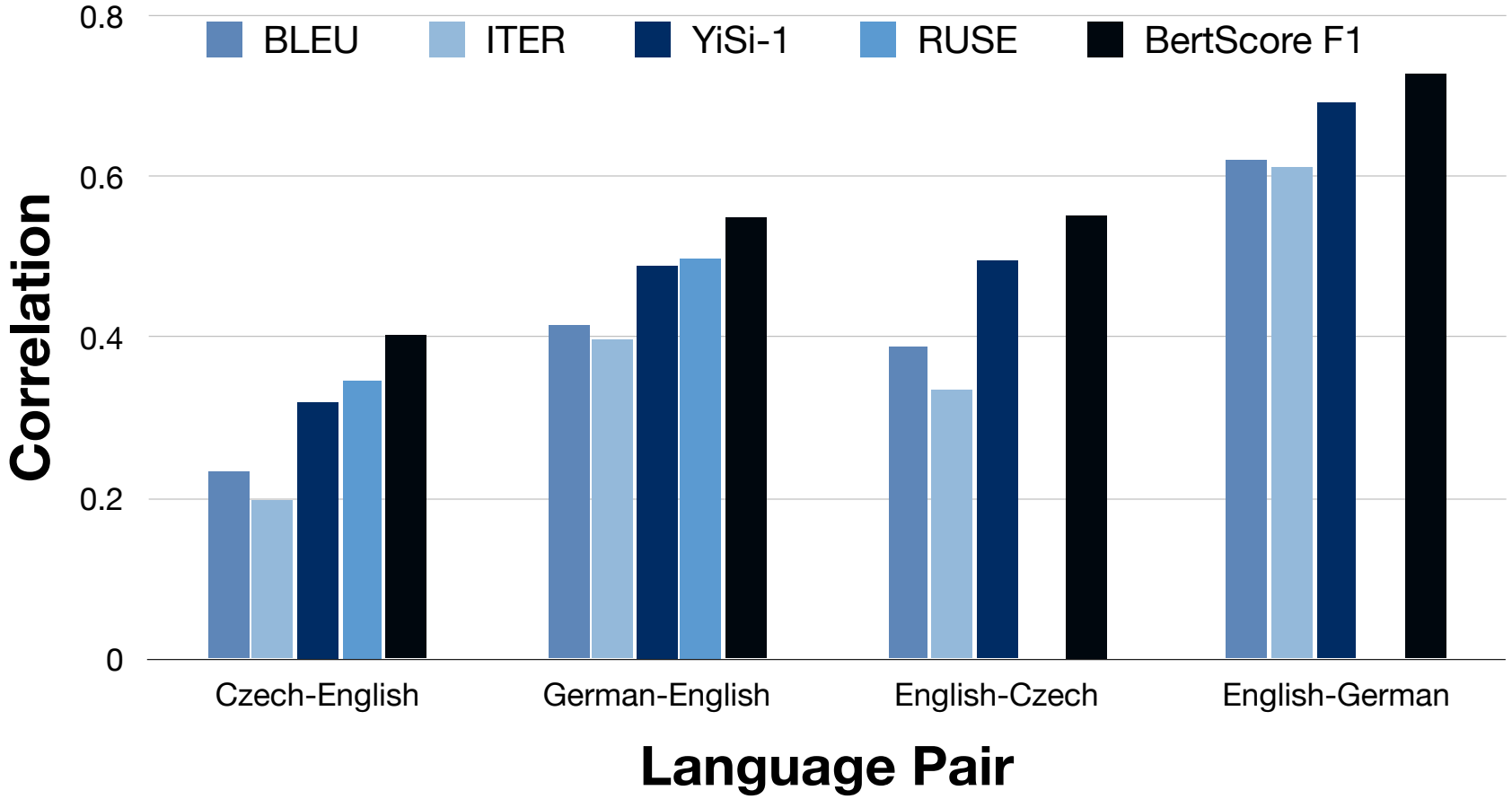
Reference: *The weather is cold today.*
Candidate: *It is freezing today.*

Reference: *The garden is nice.*
Candidate: *The garden was pretty.*

Reference: *I like apples very much.*
Candidate: *I love apples.*



Correlation Study



4 tasks

8 languages

363 systems

Download here:https://pypi.org/project/bert-score/
Or Just: pip install bert_score

Github