Natural Language Understanding, Generation, and Machine Translation

Lecture 16: Evaluating Translation and Generation

Alexandra Birch

26 February 2024 (week 6)

School of Informatics
University of Edinburgh
a.birch@ed.ac.uk

Based on slides by Adam Lopez, Rico Sennrich, Philipp Koehn

Agenda for Today

Previous lectures have surveyed all of the tools we need to *implement* a NLP system: data, effective models, and learning algorithms.

This lecture: How do we know whether what we've implemented is useable? Focus on translation, but look at generation too

Evaluation is important and difficult

Evaluation by people

Evaluation by string overlap metrics

Evaluate using Embeddings

Evaluate using metrics trained on human evaluations

Evaluation is important and difficult

Testing is crucial to good engineering

Suppose I give you a program to compute Fibonacci numbers. Suppose I give you a python interpreter. Suppose I give you a speech recognizer. Suppose I give you a self-driving car. Suppose I give you a machine translation system.

How would you decide if the implementation is correct?

What does is mean for an implementation to be correct?

Why do we need to evaluate machine translation systems?

- Decide which of two (or more) systems to use.
- Evaluate incremental changes to systems.
  - Does a new idea make it better or worse?
  - Does it change things in the intended way?
- Decide whether a system is appropriate for a given use case.
  - Understanding a restaurant menu.
  - Understanding a news about safety of a city you are visiting.
  - Translating legal notices of a product you are selling.
  - Negotiating a peace treaty.

Key questions. Are you trying to *assimilate* or *disseminate* information? Who is affected by the system, and what at are the consequences of errors for each of them?

Different translators produce different translations

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Evaluation by people

A good translation is both adequate and fluent

People can (and do) evaluate MT on many different dimensions. Two crucial ones:

Adequacy: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

Fluency: Is the output good fluent English? Is is grammatically correct? Does it use appropriate words and idioms?

Can we even measure adequacy and fluency?

Typical scales for adequacy and fluency

	Adequacy
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

	Fluency
5	flawless English
4	good English
3	non-native English
2	dis-fluent English
1	incomprehensible

Evaluate some translations

Source. Avauspelin voitto on aina tärkeä.

Reference. It is always important to win the opening match.

System 1. Victory for the game is always important.

System 2. The victory of the opening game is always important.

System 3. Victory in the opening game is always important.

Evaluate some translations

Source:

ઘટનાની જાણકારી મળતા જ ઘરે જોવાવાળાનો જમાવડ

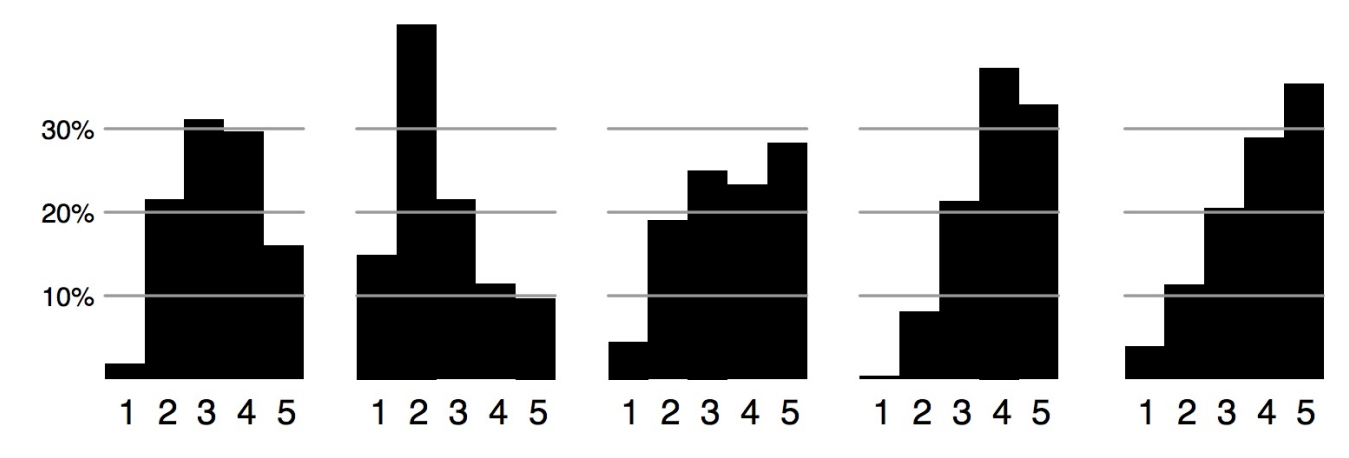
Reference: As people came to know of this, they started gathering to see this spectacle.

System 1. As soon as the incident was known, the house was flooded.

System 2. As soon as the information of the incident came to pass, there was a group of people who saw it at home.

System 3. As soon as the incident was reported, there was a meeting of people who saw it at home.

Evaluators often disagree



We can measure agreement between evaluators

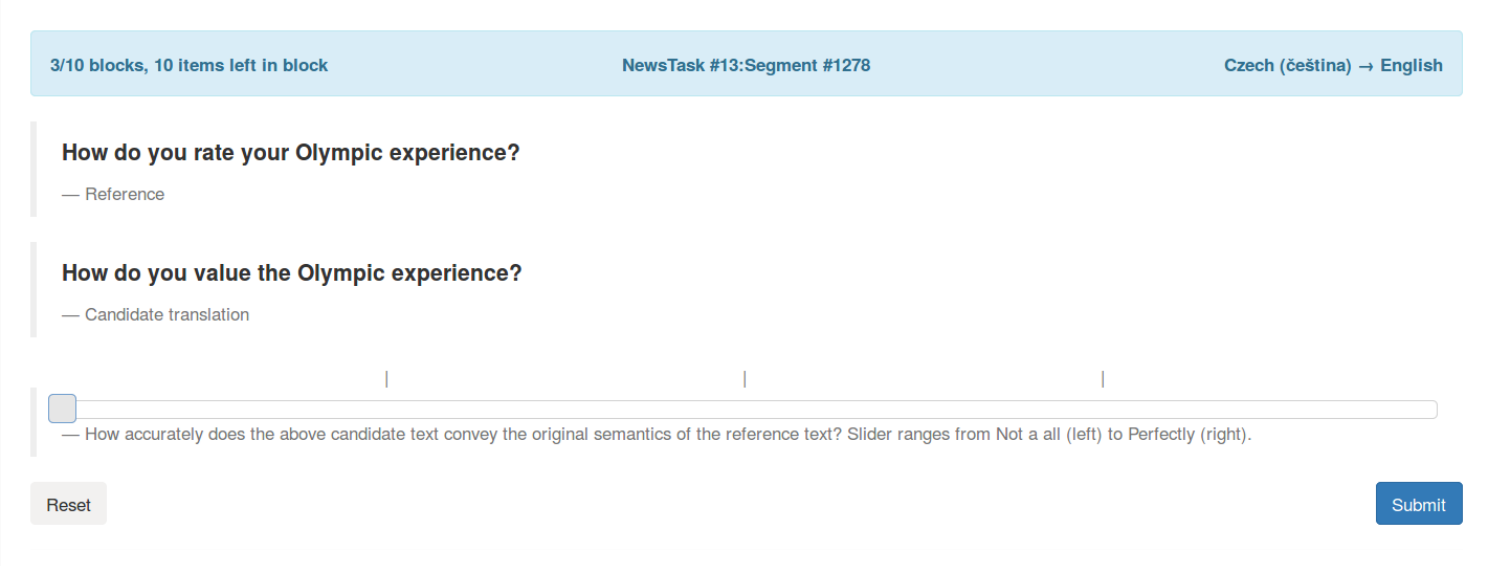
$$\kappa = \frac{p(A) - p(E)}{1 - p(E)}$$

- $p(A)$ is proportion of times that evaluators agree.
- $p(E)$ is proportion of times that they would agree by chance.

Empirically, agreement on fluency and adequacy is low, but positive. Agreement on rating (which of two translations is better?) tends to be higher, but still not very high.

Adequacy and fluency are very abstract, difficult to measure.

Direct Assessment



From Findings of the 2017 conference on machine translation (wmt17) Bojar et al. (2017)

- 100-point Likert scale allowing fine grained statistical analysis
- Normalise individual annotators, quality control with references
- Intra-annotator agreement is higher
- Continuous measurement scales in human evaluation of machine translation Graham et al. (2013)

Evaluation by string overlap metrics

Can we evaluate automatically?

A very specific use case: evaluating *incremental* changes to systems. This typically requires something automatic, due to the cost of human evaluation.

How would you decide whether to deploy a change to Google translate, which supports over 100 languages in any direction?

Idea. Human evaluators compare with a *reference translation* when they don't know the source language. We can automate this comparison.

Q. What are the pros and cons of this idea?

Idea: count all of the words that match

System 1. Victory in the opening game is always important

Reference. It is always important to win the opening match

System 2. the it opening important is match always win to

Precision. $\frac{\#\text{ of correct words}}{\#\text{ of output words}} = \frac{5}{8}$

Recall. $\frac{\#\text{ of correct words}}{\#\text{ of reference words}} = \frac{5}{9}$

F-measure. $\frac{\text{precision} \times \text{recall}}{(\text{precision}+\text{recall})/2}$

System	Precision	Recall	F-measure
System 1	.623	.55	.58
System 2	1.0	1.0	1.0

Problem. Does not account for word order.

New idea: count all of the n-grams that match

System. Victory in the opening game is always important

Reference 1. It is always important to win the opening match

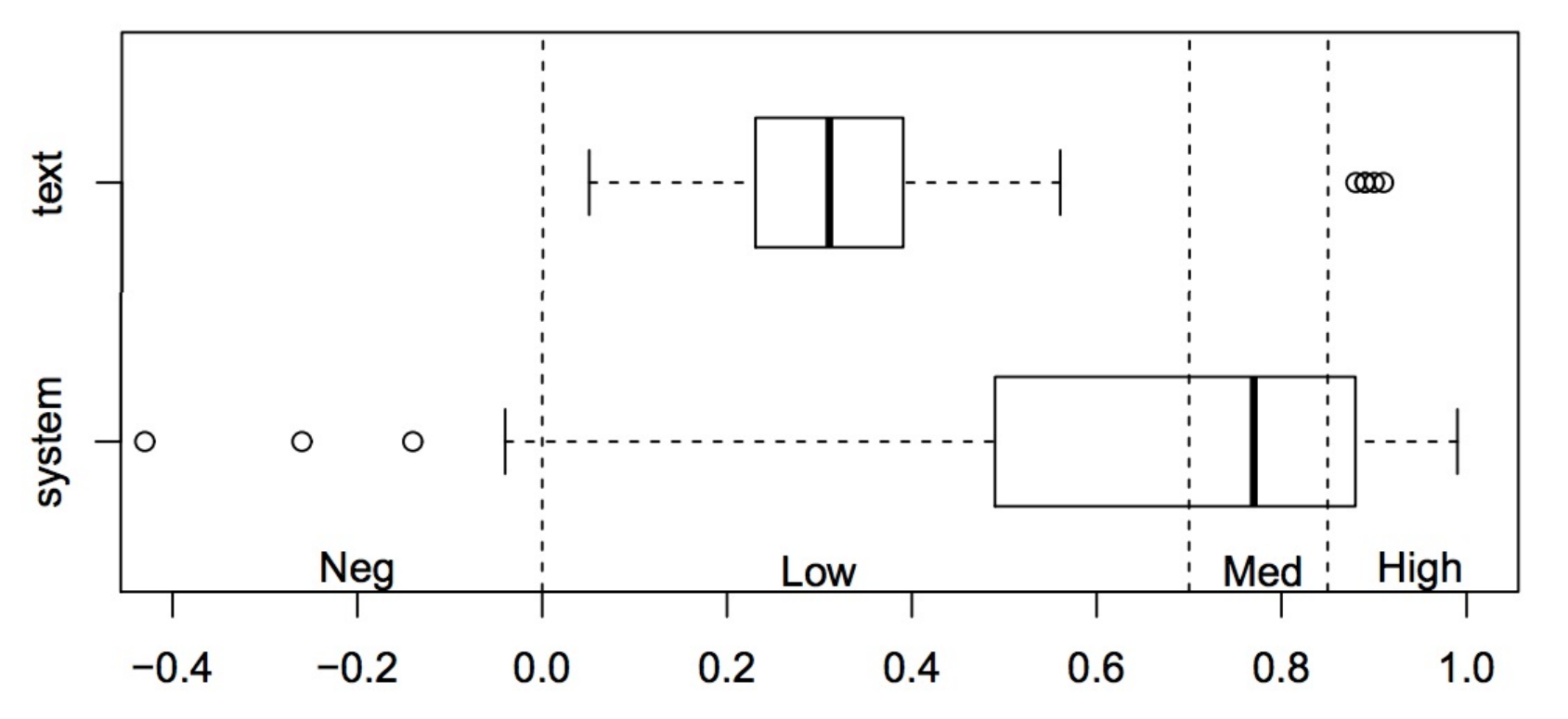
Reference 2. Opening game wins are always important.

Compute precision for n-grams of size 1 to 4 against multiple references.

Recall not well-defined in this setting. *BLEU* compares system length to an *effective reference length* and penalize if too short.

$$\text{BLEU} = \min\left(1, \frac{\text{output length}}{\text{reference length}}\right)\left(\prod_{n=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

BLEU is reasonably correlated with human ratings of MT

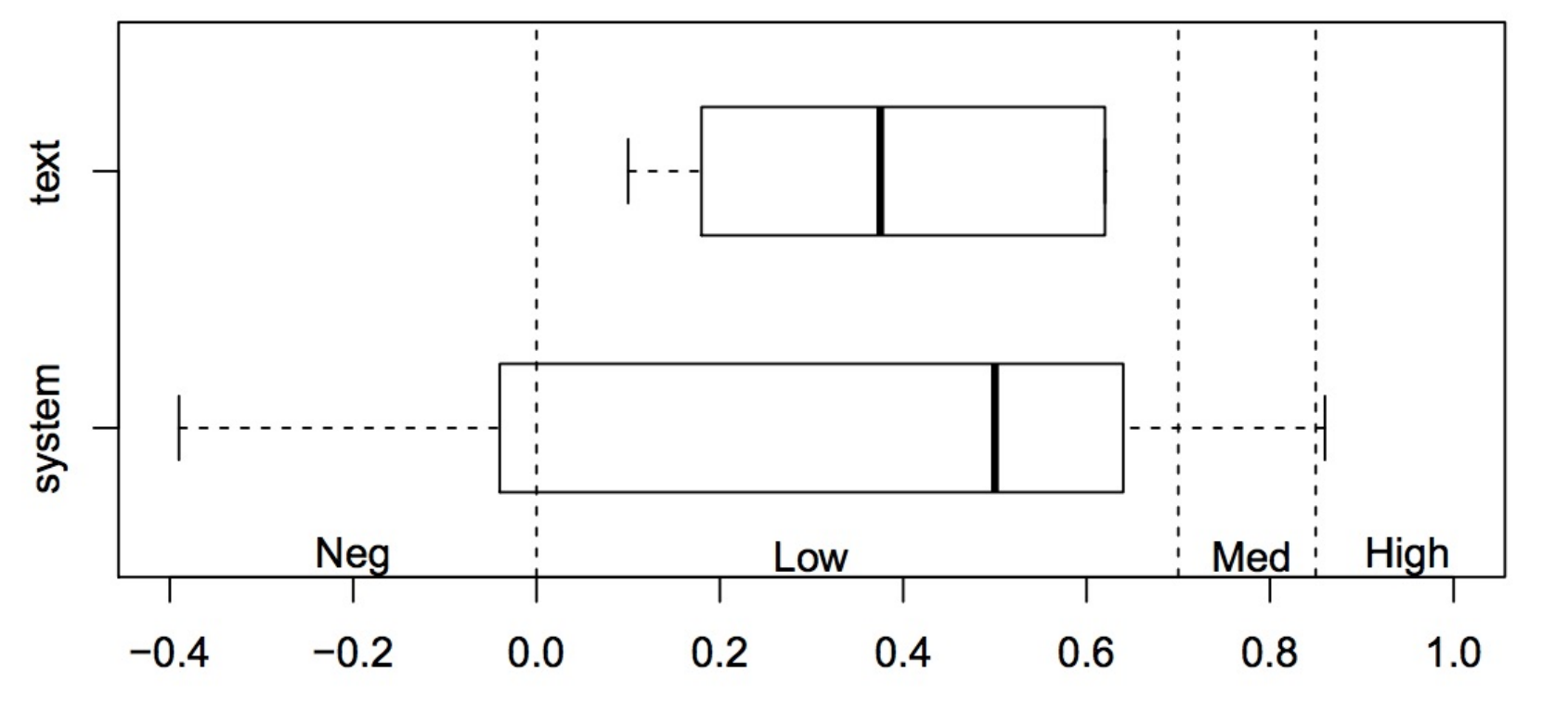


Source: Ehud Reiter, A Structured Review of the Validity of BLEU

Many studies have, over time, shown *some* correlation between BLEU and human ratings.

No studies have shown relationship to real applications.

BLEU is less correlated with human ratings of NLG



Source: Ehud Reiter, A Structured Review of the Validity of BLEU

In brief: Take BLEU with skepticism

Bleu is generally a crude measure of accuracy

- BLEU performs worse on morphologically rich languages - use character level Chrf instead
- Not all words are equally important! BLEU treats determiners and punctuation the same as names and other content words.
- BLEU is a poor proxy for both adequacy and fluency.
- BLEU isn't interpretable across datasets.
- BLEU often scores human translation low.

Evaluating Generation

- Translation: constrained by input text
- Generation: more complex task generating novel text not constrained by input text

- Evaluation is more nuanced and might need automatic metrics and human evaluation
- **Recall-Oriented Understudy for Gisting Evaluation** straightforward and granualar metric
- ROUGE-1/ROUGE-2 overlap of unigrams/bigrams between reference and summary

ROUGE-L

Longest Common Subsequence (LCS) captures sentence-level overlap

System: the entry for a big brown fox bites

Reference: the rabid fox bites Pedro

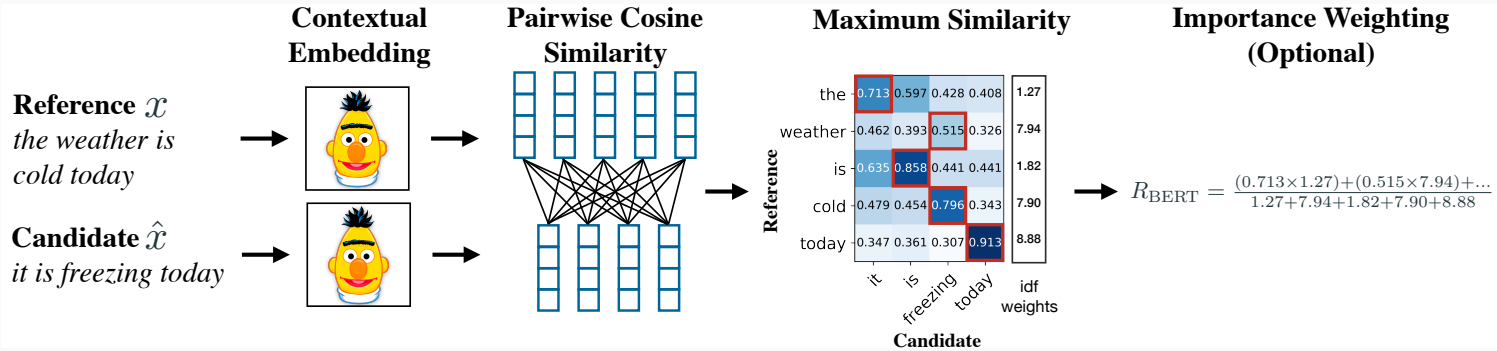
$$ROUGE - L_{recall} = \frac{LCS}{|reference|} = \frac{3}{5}$$

Evaluate using Embeddings

Embedding based metrics

- Surface level metrics: Fail to catch paraphrases, important word order differences
- Contextualized embeddings are trained to effectively capture semantic overlap, distant dependencies and ordering
- BERTScore: embeddings, pairwise cosine similarity, greedy matching, optional idf importance weighting

BERTScore



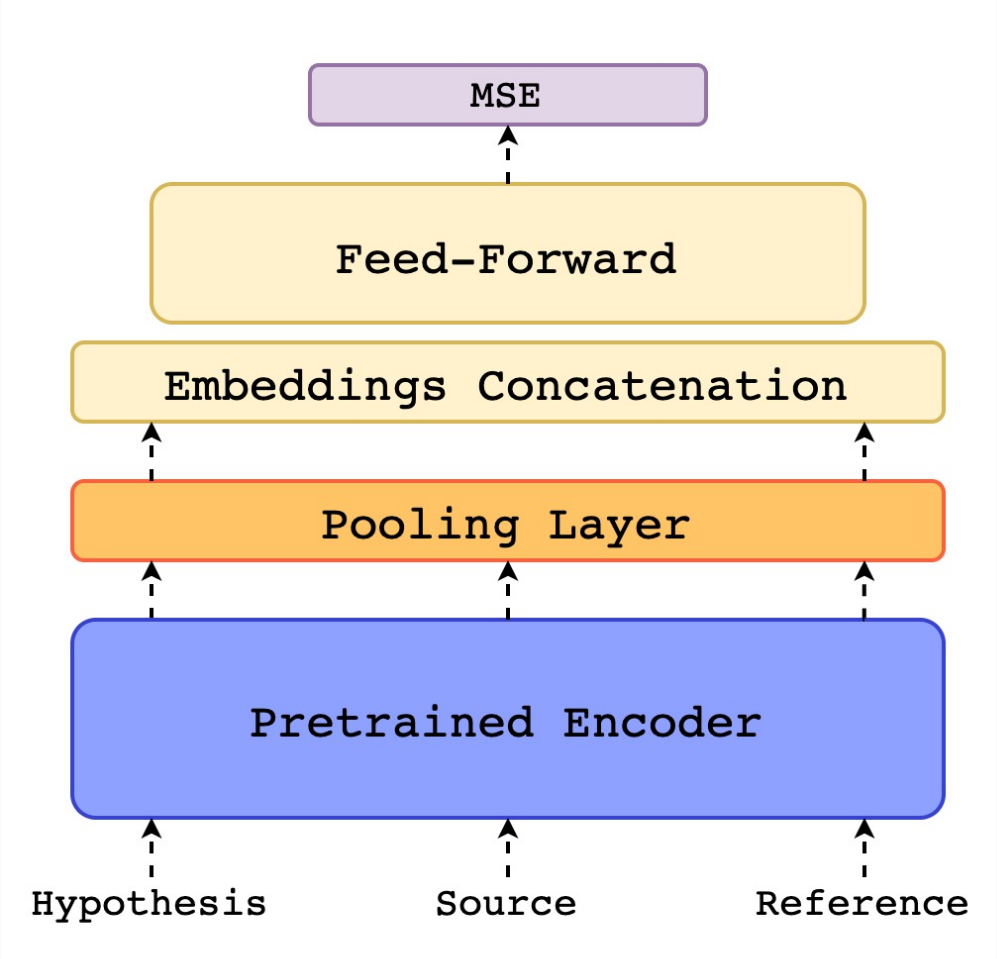
From Zhang et al. (2019)

Evaluate using metrics trained on human evaluations

Trained Metrics

- Transformers can be used to encode, decode and evaluate!
- Have 16 years of human evaluation data from WMT tasks - labelled training data
- COMET (Rei et al., 2020): Crosslingual Optimized Metric for Evaluation of Translation.
- Pretrain cross-lingual language model
- Fine-tune on human evaluations eg. Direct Assessments
- SOTA results for correlation with human judgements WMT 2019 Metrics shared task
- Trained on human evaluations for a number of language pairs - some evidence that it generalises, but certainly less reliable

COMET



From Rei et al. (2020)

Be skeptical of hype

Programmering 11 points · 4 years ago · *edited 4 years ago*

What do you believe that AI capabilities could be in the close future?

wojzaremba **OpenAI** 17 points · 4 years ago

Speech recognition and machine translation between any languages should be fully solvable.

Best Practise

- Define NLG task and objectives clearly and explicitly
- Select relevant and appropriate metrics
- Multiple metrics and methods should be used to capture different dimensions of quality
- Use a large and diverse sample of test data
- Ideally a representative and unbiased sample of human evaluators

Summary of key points (i.e. examinable content)

- Good evaluation of NLG and translation is both *really, really important* and *really, really difficult*.
- We can distinguish between two crucial concerns in MT/NLG systems: *adequacy* and *fluency*.
- Automatic evaluation metrics for iterative system development eg. BLEU score
- Typical evaluation metrics measure *n*-gram overlap with a human *reference translation*. Has many problems.
- Trained metrics correlate better with humans
- Understanding which phenomena your system handles well, and which it doesn't, requires you to *look at the data*.
- Be skeptical of claims of human-level accuracy.

Next lecture: Multilingual data for machine translation

Bibliography i

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., et al. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.