# Deep Learning for Natural Language Processing
## Evaluation of Generation Systems

UNIVERSITY OF GOTHENBURG

CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

**Richard Johansson**

richard.johansson@gu.se

# how well-defined is the generation problem?

| Machine Translation | Sentence Compression | Abstractive Summarization | Story Generation | Chitchat Dialogue |
|---|---|---|---|---|

| ⟵ | ⟶ |
|---|---|
| Less open-ended | More open-ended |
| Mostly word-level decisions | Requires more high-level decisions |
| Neural LMs more successful | **Neural LMs less successful** |
| Control is less important | **Control is more important** |
| Eval is difficult | **Eval is fiendish** |

[source]

# using human evaluators

- different protocols, mostly based on some variation of:
  - **fluency**
  - **adequacy**

**Fluency**

How do you judge the fluency of this translation?

5 = Flawless English
4 = Good English
3 = Non-native English
2 = Disfluent English
1 = Incomprehensible

**Adequacy**

How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?

5 = All
4 = Most
3 = Much
2 = Little
1 = None

(Callison-Burch et al., 2006)

- examples:

Reference: **Yesterday, stock and commodity prices fell on the world's markets.**

Output 1: Global stock markets and commodity markets fell yesterday.

Output 2: The stock market fell in Zurich.

Output 3: Around globe stock, and and also, commodities fall yesterday.

Output 4: Market and win ball rolling yesterday around electronic highly.

# automatic evaluation methods

- human judgments take too much time
- for efficient evaluation and for incremental system development, we need automatic evaluation protocols
- in most cases, they are based on various **overlap measures** between the proposed output and (one or more) references

# word error rate

| | |
|---|---|
| **Reference** | Israeli officials are responsible for airport security |
| **System** | Israeli officials responsibility of airport safety |

# word error rate

| | |
|---|---|
| **Reference** | Israeli officials are responsible for airport security |
| **System** | Israeli officials responsibility of airport safety |

# word error rate

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Reference** | Israeli | officials | are | responsible | for | airport | security |
| | | | D | S | S | | S |
| **System** | Israeli | officials | | responsibility | of | airport | safety |

# word error rate

| | |
|---|---|
| **Reference** | Israeli officials are responsible for airport security |
| | D S S S |
| **System** | Israeli officials responsibility of airport safety |

- the **word error rate** is defined as

$$WER = \frac{S + D + I}{N_{ref}} = \frac{3 + 1 + 0}{7}$$

- most commonly used in applications where there isn't much "freedom" in how to generate the output

# precision and recall at the word level

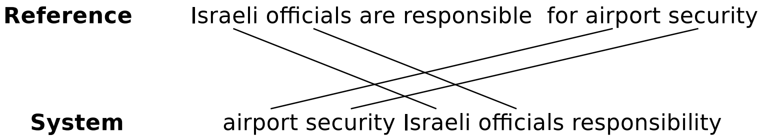
| | |
|---|---|
| **Reference** | Israeli officials are responsible for airport security |
| **System** | airport security Israeli officials responsibility |

# precision and recall at the word level

**Reference** Israeli officials are responsible for airport security

**System** airport security Israeli officials responsibility

# precision and recall at the word level




$$P = \frac{4}{5} \qquad R = \frac{4}{7}$$

- as usual, the $F$-score is the harmonic mean of $P$ and $R$
- we can also compute $P$ and $R$ for bigrams, trigrams, ...

# common metrics based on *n*-gram precision and recall

- *P* and *R* scores for *n*-grams are also called **ROUGE** scores (Lin, 2004), typically used to evaluate summarization systems
  - for instance, **ROUGE-2** *F*-score is the bigram *F*-score
- **BLEU** (Papineni et al., 2002), commonly used to evaluate machine translators, uses the precision for different *n*

# the BLEU score

| | Translation | $p_1$ | $p_2$ | $p_3$ | $p_4$ | BP |
|---|---|---|---|---|---|---|
| *Reference* | *Vinay likes programming in Python* | | | | | |
| *Sys1* | *To Vinay it like to program Python* | $\frac{2}{7}$ | 0 | 0 | 0 | 1 |
| *Sys2* | *Vinay likes Python* | $\frac{3}{3}$ | $\frac{1}{2}$ | 0 | 0 | .51 |
| *Sys3* | *Vinay likes programming in his pajamas* | $\frac{4}{6}$ | $\frac{3}{5}$ | $\frac{2}{4}$ | $\frac{1}{3}$ | 1 |

- the BLEU score uses the precision of *n*-grams of length 1–4

$$\text{BLEU} = \text{BP} \cdot \left( \prod_{i=1}^{4} p_i \right)^{\frac{1}{4}}$$

where BP is a **brevity penalty** that punishes short outputs

$$\text{BP} = \min(1, e^{1 - \frac{|R|}{|S|}})$$

# multiple references

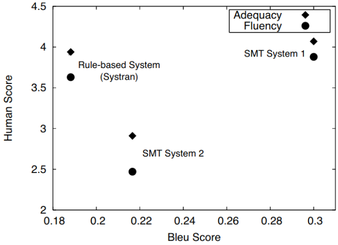
- for some tasks including MT, many possible outputs are possible
- **multiple reference outputs** are often used in evaluations

| |
|---|
| Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. |
| Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida. |
| Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida. |
| Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida. |

| |
|---|
| Appeared calm when he was taken to the American plane, which will to Miami, Florida. |

# does BLEU make sense?

- BLEU scores are reported in almost every MT paper
- but do they measure the actual quality well enough?




- generally, there tends to be a rough correlation between BLEU and human scores
- Callison-Burch et al. (2006) claim that BLEU might be misleading when comparing systems **of different types**
- METEOR (Banerjee and Lavie, 2005) addresses some of the word matching issues with BLEU

# some implementations

- **SacreBLEU** (Post, 2018) is a standardized BLEU implementation in Python https://github.com/mjpost/sacreBLEU/
- **ROUGE** 2.0: http://rxnlp.com/rouge-2-0
- **METEOR**: https://www.cs.cmu.edu/~alavie/METEOR/

# references


S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshops*.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *EACL*.

C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL workshops*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

M. Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*.