Applications

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
RL FOR MACHINE TRANSLATION

Slides adapted from Phillip Koehn

Evaluation

- How good is a given machine translation system?
- Hard problem, since many different translations acceptable
  ⟶ semantic equivalence / similarity
- Evaluation metrics
  - subjective judgments by human evaluators
  - automatic evaluation metrics
  - task-based evaluation, e.g.:
    - how much post-editing effort?
    - does information come across?

Ten Translations of a Chinese Sentence

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.

(a typical example from the 2001 NIST evaluation set)

Adequacy and Fluency

- Human judgement
  - given: machine translation output
  - given: source and/or reference translation
  - task: assess the quality of the machine translation output
- Metrics

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?
This involves both grammatical correctness and idiomatic word choices.

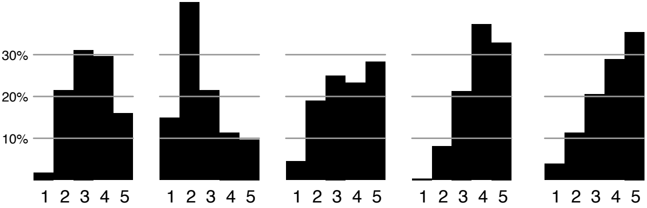
Fluency and Adequacy: Scales

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	disfluent English
1	incomprehensible

Evaluators Disagree

- Histogram of adequacy judgments by different human evaluators



(from WMT 2006 evaluation)

Goals for Evaluation Metrics

Low cost: reduce time and money spent on carrying out evaluation

Tunable: automatically optimize system performance towards metric

Meaningful: score should give intuitive interpretation of translation quality

Consistent: repeated use of metric should give same results

Correct: metric must rank better systems higher

Other Evaluation Criteria

When deploying systems, considerations go beyond quality of translations

Speed: we prefer faster machine translation systems

Size: fits into memory of available machines (e.g., handheld devices)

Integration: can be integrated into existing workflow

Customization: can be adapted to user's needs

Automatic Evaluation Metrics

- Goal: computer program that computes the quality of translations
- Advantages: low cost, tunable, consistent
- Basic strategy
  - given: machine translation output
  - given: human reference translation
  - task: compute similarity between them

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

- Precision

$$\frac{\underline{\textbf{correct}}}{\underline{\textbf{output-length}}} = \frac{3}{6} = 50\%$$

- Recall

$$\frac{\underline{\textbf{correct}}}{\underline{\textbf{reference-length}}} = \frac{3}{7} = 43\%$$

- F-measure

$$\frac{\underline{\textbf{precision}} \times \underline{\textbf{recall}}}{(\underline{\textbf{precision}} + \underline{\textbf{recall}})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

Metric	**System A**	**System B**
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

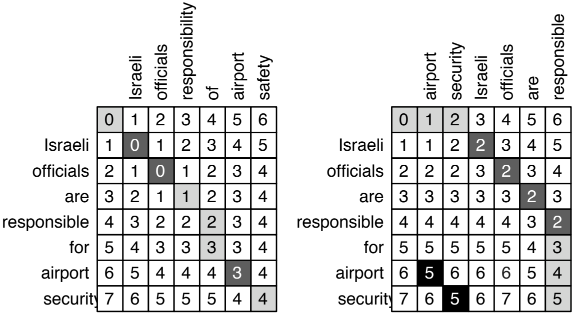
flaw: no penalty for reordering

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

$$\text{wer} = \frac{\underline{\textbf{substitutions}} + \underline{\textbf{insertions}} + \underline{\textbf{deletions}}}{\underline{\textbf{reference-length}}}$$

Example



		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
securit	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
securit	7	6	5	6	7	6	5

Metric	**System A**	**System B**
word error rate (wer)	57%	71%

BLEU

- *N*-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{bleu} = \min\left(1, \frac{\underline{\textbf{output-length}}}{\underline{\textbf{reference-length}}}\right)\left(\prod_{i=1}^{4} \underline{\textbf{precision}}_i\right)^{\frac{1}{4}}$$

- Typically computed over the entire corpus, not single sentences

Example

SYSTEM A: Israeli officials responsibility of airport safety

2-GRAM MATCH (Israeli officials) 1-GRAM MATCH (airport)

REFERENCE: Israeli officials are responsible for airport security

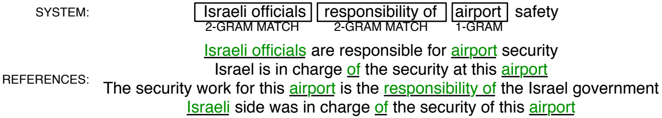
SYSTEM B: airport security Israeli officials are responsible

2-GRAM MATCH (airport security) 4-GRAM MATCH (Israeli officials are responsible)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
bleu	0%	52%

Multiple Reference Translations

- To account for variability, use multiple reference translations
  - n-grams may match in any of the references
  - closest reference length used
- Example



SYSTEM: Israeli officials responsibility of airport safety

2-GRAM MATCH 2-GRAM MATCH 1-GRAM

REFERENCES:

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

Challenge

- Most machine learning approaches tune on likelihood
- How can we measure BLEU (or other metrics)
- And how does this work with decoding

Challenge

- Most machine learning approaches tune on likelihood
- How can we measure BLEU (or other metrics)
- And how does this work with decoding . . . reinforcement learning