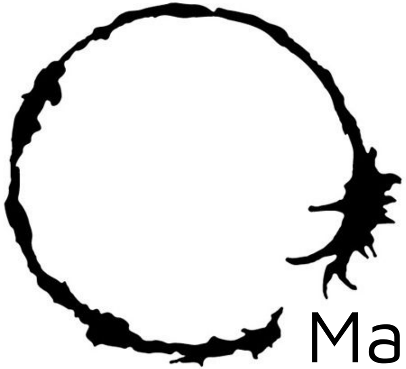
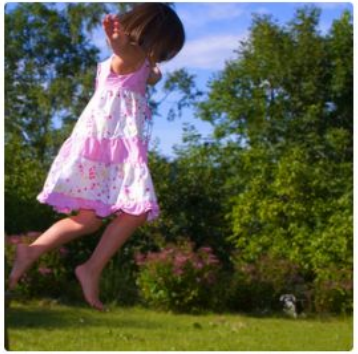


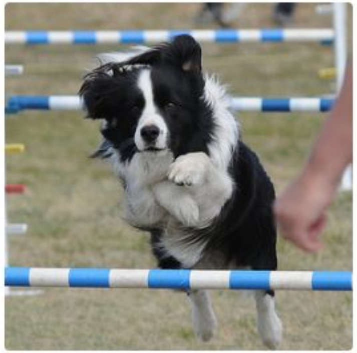
Machine Translation Evaluation

Onur Aydın

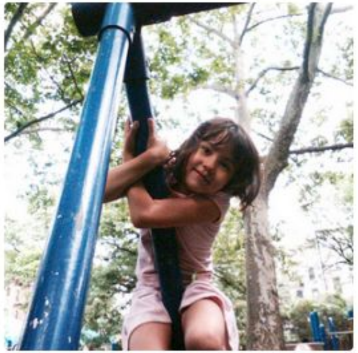
Natural Language Processing



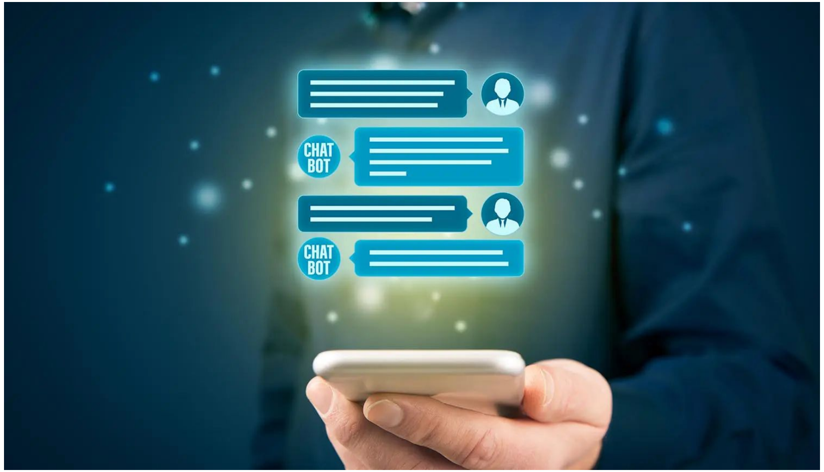
"girl in pink dress is jumping in air."



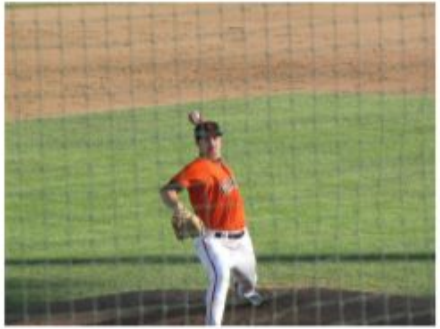
"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



No unique solution



- A baseball winds up to pitch the ball.
- A pitcher throwing the ball in a baseball game.
- A pitcher throwing a baseball on the mound.
- A baseball player pitching a ball on the mound.
- A left-handed pitcher throwing for the San Francisco giants.

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.

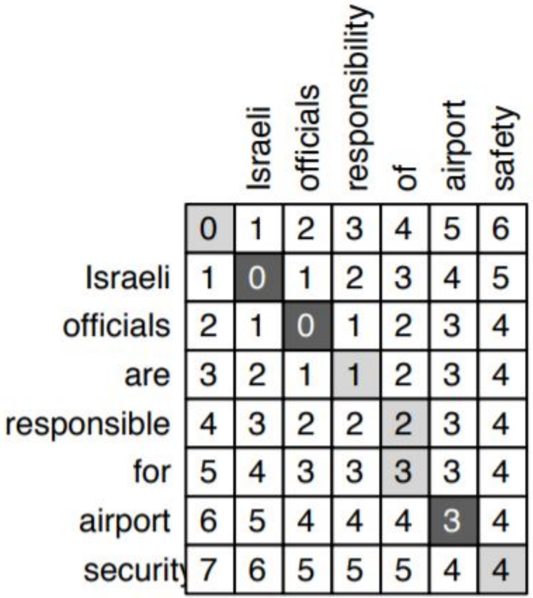
A Machine Translation metric must be...

- Quick
- Inexpensive
- Language-independent
- Correlated with human evaluation

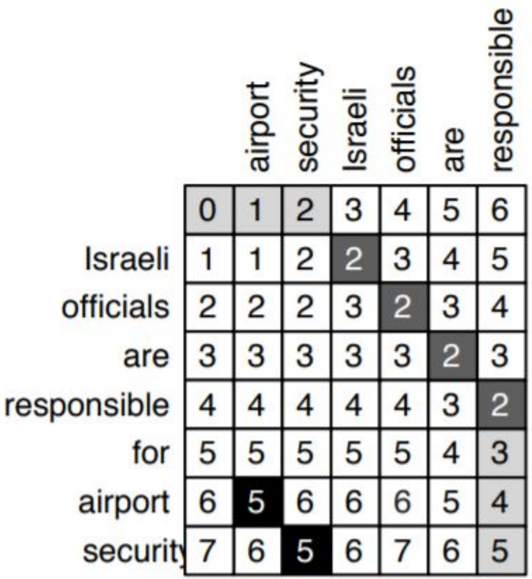
Agenda

- Word Error Rate
- Bilingual Evaluation Understudy (BLEU) [Link]
- Translation Edit Rate (TER) [Link]
- Metric for Evaluation of Translation with Explicit Ordering (METEOR) [Link]

Word Error Rate



		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%

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference-length}}$$

BLEU Score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BLEU Score

$$\text{B}{\scriptstyle\text{LEU}} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

modified N-gram precisions

BLEU Score

Modified Precisions

Reference: the cat is on the mat

Machine Translation: the the the the the the the

Precisions = 7/7 = 1

Modified Precisions = 2/7

BLEU Score

Modified Precisions

Reference: the cat is on the mat

Machine Translation: the the the the the the the

Precisions = 7/7 = 1

Modified Precisions = 2/7

Modified Precisions on Bigrams

Reference-1: the cat is on the mat

Reference-2: there is a cat on the mat

Machine Translation: the cat the cat on the mat

the cat	2	1
cat the	1	0
cat on	1	1
on the	1	1
the mat	1	1

Modified Precisions = 4/6

BLEU Score

$N = 4$

$w_n = 1/N$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

modified N-gram precisions

$$\text{Brevity Penalty (BP)} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c: length of candidate translation

r: length of reference translation

BLEU Score

$$\text{B}\textsc{leu}= \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\log \text{B}\textsc{leu} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

TER Score

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

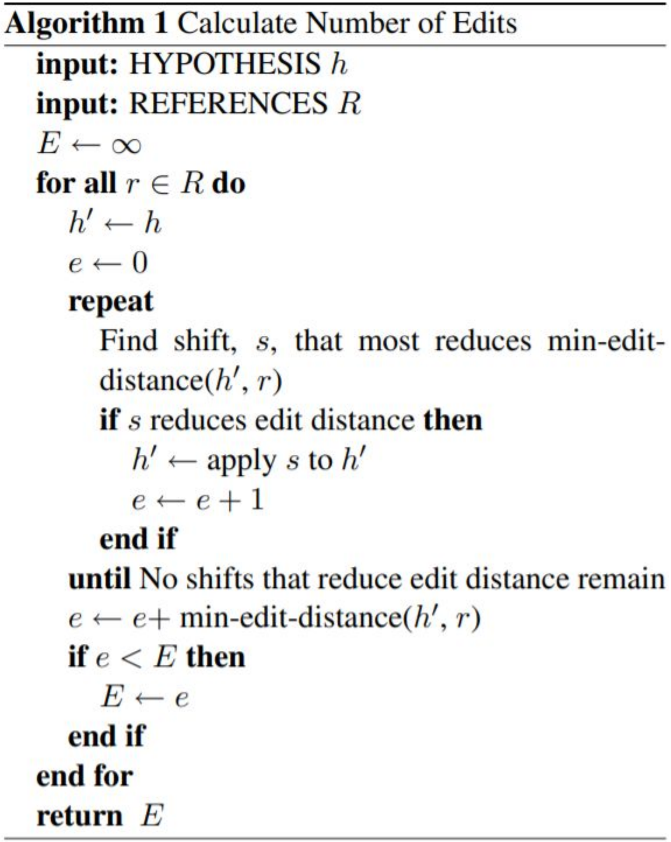
Possible edits:

- Insertion
- Deletion
- Substitution of single words

Dynamic Programming

- Shifts of word sequences

Greedy Search



Algorithm 1 Calculate Number of Edits

```
input: HYPOTHESIS $h$
input: REFERENCES $R$
$E \leftarrow \infty$
for all $r \in R$ do
    $h' \leftarrow h$
    $e \leftarrow 0$
    repeat
        Find shift, $s$, that most reduces min-edit-distance($h'$, $r$)
        if $s$ reduces edit distance then
            $h' \leftarrow$ apply $s$ to $h'$
            $e \leftarrow e + 1$
        end if
    until No shifts that reduce edit distance remain
    $e \leftarrow e +$ min-edit-distance($h'$, $r$)
    if $e < E$ then
        $E \leftarrow e$
    end if
end for
return $E$
```

TER Score

REF: saudi arabia denied this week information published in the american new york times

TRA: this week the saudis denied information published in the new york times

THIS WEEK THE SAUDIS denied information published in the new york times

THE SAUDIS denied **THIS WEEK** information published in the new york times [SHIFT]

THE **SAUDIS** denied THIS WEEK information published in the **AMERICAN** new york times [INSERTION]

THE ARABIA denied THIS WEEK information published in the AMERICAN new york times [SUBSTITUTION]

SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times [SUBSTITUTION]

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}} = 4/13$$

METEOR Score

$$Score = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

METEOR Score

$$Score = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

P: Unigram Precision

R: Unigram Recall

* recall weighted 9 times more than precision

- Exact
- Porter stem
- WN Stem
- WN Synonymy

METEOR Score

$$Score = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

REF: “the president spoke to the audience”

TRA: “the president then spoke to the audience”

#chunks = 2 > “the president then spoke to the audience”

#unigrams_matched = 6

METEOR Score

$$Score = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

Comparison

	BLEU	TER	METEOR
Year	2002	2006	2007
Precision	✓	✓	✓
Recall	X	X	✓
N-gram	1, 2, 3, 4	-	1
Synonym	X	X	✓

References

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). **BLEU: a method for automatic evaluation of machine translation.** In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006, August). **A study of translation edit rate with targeted human annotation.** In Proceedings of association for machine translation in the Americas (Vol. 200, No. 6).

Lavie, A., & Agarwal, A. (2007, June). **METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments.** In Proceedings of the second workshop on statistical machine translation (pp. 228-231).