Evaluation

Machine Translation

Samuel Läubli, Mathias Müller

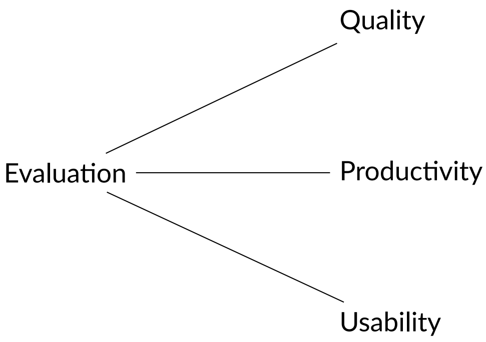
Institute of Computational Linguistics
University of Zurich

February 26, 2019

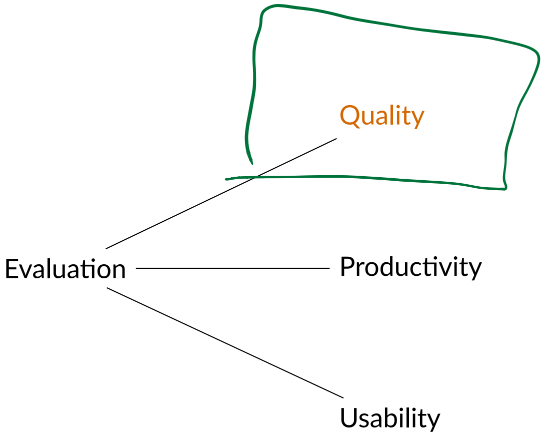
University of Zurich[UZH]

Topics

What are we evaluating?



What are we evaluating?



Example sentence

The world is a stage, but the play is badly cast.

– Oscar Wilde

Evaluation of quality: requirements

metric

A metric that evaluates translation quality should meet the following criteria:

Evaluation of quality: requirements

metric

7 8

A metric that evaluates translation quality should meet the following criteria:

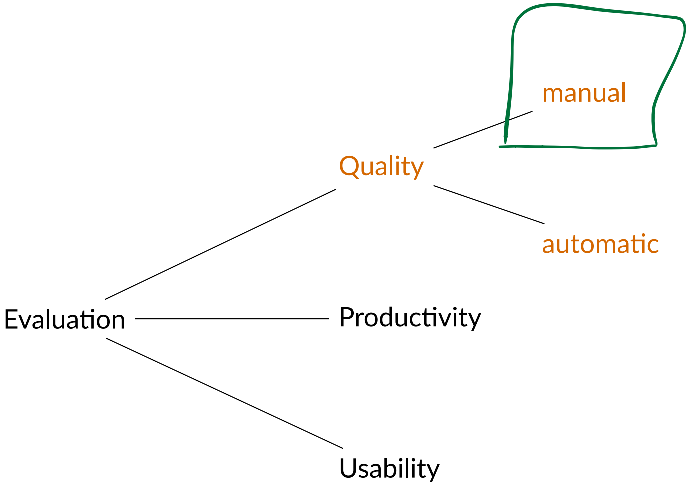
- **low cost**: evaluation should be fast and cheap
- **compelling**: metric should be easy to interpret
- **consistent**: repeated evaluations should lead to the same results
- **correct**: evaluation should be truthful.

Evaluation of quality: requirements

A metric that evaluates translation quality should meet the following criteria:

- **low cost**: evaluation should be fast and cheap
- **compelling**: metric should be easy to interpret
- **consistent**: repeated evaluations should lead to the same results
- **correct**: evaluation should be truthful. $\rightarrow$ Problem: Subjectivity. There is no (singular) «thruth» (ground truth) in translation.

How to evaluate quality?



Pros and Cons

Manual evaluation

+ more reliable
- costly
- slow

Automatic evaluation

- less reliable
+ cheap
+ fast

Topics

1. Introduction
2. Manual Evaluation
3. Automatic Evaluation
4. Summary

Topics

Example

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

Example

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

How would <u>you</u> ask people to evaluate?

bit.ly/2TOGDeK

Example

Idea 1

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

On a scale from 1 to 5,

- how **adequate** is the translation? (sentence still has the same meaning)
- how **fluent** is the translation? (grammatical, suitable style)

Example

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

DeepL:
Die Welt ist eine Bühne, aber das Stück ist schlecht besetzt.

Example

Idea 2

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

DeepL:
Die Welt ist eine Bühne, aber das Stück ist schlecht besetzt.

Which translation is better?

- Google Translate > DeepL
- Google Translate = DeepL
- Google Translate < DeepL

Absolute manual evaluation

absolute

Machine-translated sentences can be evaluated with absolute numbers. As a convention, we evaluate **adequacy** and **fluency** on a five point Likert scale.

Absolute manual evaluation

1 – 5

Machine-translated sentences can be evaluated with absolute numbers. As a convention, we evaluate **adequacy** and **fluency** on a five point Likert scale.

→ What does a fluency of 4 mean exactly?

Absolute manual evaluation: example (WMT 2006)

WMT

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	○ ○ ○ ○ ◉ 1 2 3 4 5	○ ○ ○ ○ ◉ 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	○ ○ ◉ ○ ○ 1 2 3 4 5	○ ○ ◉ ○ ○ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	○ ○ ○ ◉ ○ 1 2 3 4 5	○ ○ ○ ◉ ○ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	○ ○ ◉ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ◉ 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	○ ○ ◉ ○ ○ 1 2 3 4 5	○ ○ ◉ ○ ○ 1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Source: Koehn and Monz, 2006

Absolute manual evaluation

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

Source: Koehn and Monz, 2006

Absolute manual evaluation

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

→ What is the difference between «much meaning» and «most meaning»?

Source: Koehn and Monz, 2006

Absolute manual evaluation: problems

Problems

- unclear definitions
- different people assign different scores on average
- sometimes, annotators cannot reproduce their own evaluation
- evaluation of adequacy and fluency is highly correlated – hard to tell apart

Relative manual evaluation

ranking

Evaluations are generally more consistent if two or more systems are compared, instead of given absolute scores

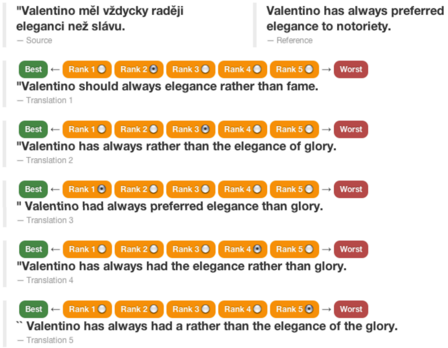
Relative manual evaluation: example (WMT 2013)

For each ranking task, the judge is presented with a source segment, a reference translation, and the outputs of five systems (anonymized and randomly-ordered). The following simple instructions are provided:

You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).

Quelle: Bojar et al., 2013

Relative manual evaluation: example (WMT 2013)



Source: Bojar et al., 2013

Relative manual evaluation: Pairwise Ranking

Relative evaluations result in pair-wise relationships between systems A, B:

A better than B	tie	B better than A
41	12	59

Relative manual evaluation: Pairwise Ranking

Relative evaluations result in pair-wise relationships between systems A, B:

A better than B	tie	B better than A
41	12	59

$\rightarrow$ Is system A truly better than system B, or are differences due to chance?

Relative manual evaluation: Pairwise Ranking – Significance

Null hypothesis: Quality gap between systems A and B due to random variation.

Alternative hypothesis: Quality gap between systems A and B not due to chance.

To reject the null hypothesis, we expect

- less than 5% probability that difference is due to random variation $\rightarrow$ difference statistically significant at 95% ($p < 0.05$)

or, to be even more strict,

- less than 1% probability that difference is due to random variation $\rightarrow$ difference statistically significant at 99% ($p < 0.01$)

Relative manual evaluation: Pairwise Ranking – Significance

Statistical significance can be tested with a *sign test*.
Example in R:

```
> binom.test(59, 100, p=0.5, alternative="two.sided")

  Exact binomial test

data: 59 and 100
number of successes = 59, number of trials = 100,
p-value = 0.08863
alternative hypothesis: true probability of success is
not equal to 0.5
```

...

Relative manual evaluation: Pairwise Ranking – Significance

Relative evaluations result in pair-wise relationships between systems A, B:

A better than B	tie	B better than A
41	12	59

$\rightarrow$ Is system A truly better than system B, or are differences due to chance?

Relative manual evaluation: Pairwise Ranking – Significance

Relative evaluations result in pair-wise relationships between systems A, B:

FAIL ? SUCCESS

A better than B	tie	B better than A
41	12	59

→ Is system A truly better than system B, or are differences due to chance?
→ Difference in quality is not statistically significant, i.e. random.

Topics

Data

Data

Our complete data is split into three parts: a training set, a validation set and a test set. Rules:

- Size of test set: 1000 to 2000 sentences
- select those sentences at random!
- automatic evaluation during development of a system
- manual evaluation before deployment of a system

Intuition

How do we evaluate translations automatically?

Any method for automatic evaluation is a function σ that computes the similarity between a machine translated segment («hypothesis») *h* and 1 or more reference translations *r*

$$\text{score} = \sigma(h, r) \tag{1}$$

0% 100%

Similarity measure usually between 0.0 and 1.0, or ~~0 and 00 %~~.

Prerequisites

1 Daten (Testset)

2 $\sigma(h, r)$

- Similarity function σ («metric»)
- 1..n reference translations for each sentence to be evaluated

3 h

4 r

Precision, Recall, F-Measure

- **Precision** = $\frac{\text{correct}}{\text{hyp length}}$
  How many words in the hypothesis are in the reference translation?
- **Recall** = $\frac{\text{correct}}{\text{ref length}}$
  How many words in the reference translation are in the hypothesis?
- **F1-Measure** = $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
  Harmonic mean of precision and recall.

Precision, Recall, F-Measure: Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{correct}}{\text{hyp length}} =$$

$$\text{Recall} = \frac{\text{correct}}{\text{ref length}} =$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} =$$

Precision, Recall, F-Measure: Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{correct}}{\text{hyp length}} = \frac{3}{6} = 0.5 = 50.0\,\%$$

$$\text{Recall} = \frac{\text{correct}}{\text{ref length}} =$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} =$$

Precision, Recall, F-Measure: Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{correct}}{\text{hyp length}} = \frac{3}{6} = 0.5 = 50.0\,\%$$

$$\text{Recall} = \frac{\text{correct}}{\text{ref length}} = \frac{3}{7} = 0.429 = 42.9\,\%$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} =$$

Precision, Recall, F-Measure: Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

Precision $= \frac{\text{correct}}{\text{hyp length}} = \frac{3}{6} = 0.5 = 50.0\,\%$

Recall $= \frac{\text{correct}}{\text{ref length}} = \frac{3}{7} = 0.429 = 42.9\,\%$

F1-Measure $= 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} = 2 \cdot \frac{0.5\cdot 0.429}{0.5+0.429} = 2 \cdot \frac{0.214}{0.929} = 0.461 = 46.1\,\%$

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision =

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision = $100.0\,\%$

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision = $100.0\,\%$ $\rightarrow$ word order does not matter

Word Error Rate (WER)

WER

Minimal edit distance (Levenshtein distance) of hypothesis to reference translation:

$$\text{WER} = \frac{\text{min(substitutions + insertions + deletions)}}{\text{ref length}}$$

hypothesis $\longrightarrow$ reference

Word Error Rate (WER): Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

$$\text{WER} = \frac{\text{min(substitutions + insertions + deletions)}}{\text{ref length}} =$$

Word Error Rate (WER): Example

Hypothesis:
Israeli officials responsibility of airport safety

Reference:
Israeli officials are responsible for airport security

$$\text{WER} = \frac{\text{min(substitutions + insertions + deletions)}}{\text{ref length}} = \frac{4}{7} = 0.571 = 57.1\,\%$$

Word Error Rate (WER): Problem

Hypothesis:
This airport's security is the responsibility of the Israeli security officials

Reference:
Israeli officials are responsible for airport security

Word Error Rate (WER): Problem

Hypothesis:
This airport's security is the responsibility of the Israeli security officials

Reference:
Israeli officials are responsible for airport security

WER >$100\,\%$

Word Error Rate (WER): Problem

Hypothesis:
This airport's security is the responsibility of the Israeli security officials

Reference:
Israeli officials are responsible for airport security

WER $>100\,\%$ $\rightarrow$ cares too much about exact sequence of words in the reference

Translation Error Rate[1] (TER)

TER

TER (Snover et al., 2006) is WER with a twist: moving an entire phrase (phrasal shift) counts as 1 edit operation.

[1] Also known as Translation Edit Rate.

Bilingual Evaluation Understudy (BLEU)

BLEU (Papineni et al., 2002) is by far the most popular evaluation metric for translation quality. Core ideas:

- compute ngram overlap of the hypothesis with multiple reference translations[1]
- No recall; compensated with a **«Brevity Penalty»** 1
- final value is a weighted geometric mean of **ngram precision** (usually n=1,2,3,4). 2
- computed for a corpus, not a single sentence, otherwise ngram precision for high orders (e.g. n=4) would be 0 most of the time

[1]Actually, we often use only one reference.

BLEU: Brevity Penalty

$$\text{BP} = \text{min}\left(1.0, \text{exp}\left(1 - \frac{\text{ref length}}{\text{hyp length}}\right)\right)$$

- «punish» if hypothesis is shorter than reference
- multiple references: use the length of the reference that is closest to hypothesis length (s. Koehn, 2010, S. 227)

BLEU: Brevity Penalty

$$\text{BP} = \min\left(1.0, \exp\left(1 - \frac{\text{ref length}}{\text{hyp length}}\right)\right)$$

```
In [175]: brevity_penalty(hyp_length=5., ref_length=5.)
Out[175]: 1.0

In [176]: brevity_penalty(hyp_length=5., ref_length=6.)
Out[176]: 0.8187307530779819

In [177]: brevity_penalty(hyp_length=5., ref_length=7.)
Out[177]: 0.6703200460356393

In [178]: brevity_penalty(hyp_length=5., ref_length=100.)
Out[178]: 5.602796437537268e-09

In [179]: brevity_penalty(hyp_length=6., ref_length=5.)
Out[179]: 1.0

In [180]: brevity_penalty(hyp_length=7., ref_length=5.)
Out[180]: 1.0
```

BLEU: ngram precision

$N = 4$

2

$$P = \left(\prod_{n=1}^{N} \lambda_n p_n \right)^{\frac{1}{N}}$$

$\lambda_1 = 1.0$

$\lambda_2 = 1.0$

$\lambda_3 = 1.0$

$\lambda_4 = 1.0$

- N: highest ngram order (usually 4)
- n: ngram precision of ngram order n
- λ_n: weight of ngram precision of order n (usually 1.0)

BLEU

$$\text{BLEU} = \overset{1}{\text{BP}} \cdot \overset{2}{\text{P}}$$

$$= \underbrace{\min\left(1.0, \exp\left(1 - \frac{\text{ref-länge}}{\text{hyp-länge}}\right)\right)}_{\text{BP}} \cdot \underbrace{\left(\prod_{n=1}^{N} \lambda_n p_n\right)^{\frac{1}{N}}}_{\text{ngram precision}}$$

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams:

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible)

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams:

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible)

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 =$ 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 =$ 4/5

3-grams:

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2$ = 4/5

3-grams: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible)

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2$ = 4/5

3-grams: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3$ = 2/4

4-grams:

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2$ = 4/5

3-grams: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3$ = 2/4

4-grams: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible)

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2$ = 4/5

3-grams: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3$ = 2/4

4-grams: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible) $\rightarrow p_4$ = 1/3

Brevity Penalty:

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Refrence:
Israeli officials are responsible for airport security

1-grams: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1$ = 6/6

2-grams: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2$ = 4/5

3-grams: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3$ = 2/4

4-grams: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible) $\rightarrow p_4$ = 1/3

Brevity Penalty: $\min\left(1.0, \exp\left(1 - \frac{7}{6}\right)\right) = 0.846$

BLEU: Example

Hypothesis:
airport security Israeli officials are responsible

Reference:
Israeli officials are responsible for airport security

$$
\begin{aligned}
\text{BLEU} &= BP \cdot (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}} \\
&= 0.846 \cdot \left(\frac{6}{6} \cdot \frac{4}{5} \cdot \frac{2}{4} \cdot \frac{1}{3}\right)^{\frac{1}{4}} \\
&= 0.511 \\
(&= \text{often reported as 51.1, as percent value.})
\end{aligned}
$$

BLEU: Several References

For several references,

- an n-gram is covered if it appears in *any* reference (but note clipping)
- brevity penalty is
  - the one reference length that is closest to the hypothesis length
  - or the shorter length, if two references (e.g. 9, 11) have the same distance to hypothesis length (e.g. 10)

BLEU: Clipping

Hypothesis:
are are are are are are are

Reference:
Israeli officials are responsible for airport security

every ngram counts as correct *only* as often as it appears in the reference

BLEU: Clipping

Clipping

Hypothesis:
are are are are are are are

Reference:
Israeli officials are responsible for airport security

every ngram counts as correct *only* as often as it appears in the reference

$\rightarrow$ 1-gram precision is 1/7, instead of 7/7!

BLEU: Clipping – Example

Hypothesis:
the the the the the the the

Reference 1:
the cat is on the mat

Reference 2:
there is a cat on the mat

1-gram precision p_1 =
2-gram precision p_2 =

BLEU: Clipping – Example

Hypothesis:
the the the the the the the

Reference 1:
the cat is on the mat

Reference 2:
there is a cat on the mat

1-gram precision p_1 = 2/7
2-gram precision p_2 =

BLEU: Clipping – Example

Hypothesis:
the the the the the the the

Reference 1:
the cat is on the mat

Reference 2:
there is a cat on the mat

1-gram precision p_1 = 2/7
2-gram precision p_2 = 0/7

BLEU: Problems

- **Ignores relevance of words**

  Some words are vital in a translation, others unimportant; with BLEU all have the same weight

  - Example:
  - Reference: «gave it to Trump»
  - Hypothesis «gave it at Trump» gets a worse score than «gave it to rhododendron»

- **BLEU value is very context-dependent**

  value depends on things like number of references, language, domain, preprocessing steps such as tokenisation etc.

- **As MT gets better, BLEU becomes more inadequate**

  Is BLEU still the way to go for NMT?

---

see also Callison-Burch et al., 2006

METEOR

METEOR (Banerjee and Lavie, 2005) is a popular alternative (or complementary) to BLEU

- idea: recall is more important than precision to make sure meaning is covered in the translation
- Alignment of words in hypothesis and reference
- 3-step matching:
  - **surface form**; or else
  - **stem** (via stemming) with penalty; or else
  - **semantic class** (via Wordnet) with penalty; or else
  - no matching possible

METEOR: Problems

- many hyperparameters (e.g.. weights for stem and synonym matches)
- more complicated computation than BLEU
- language-dependent: needs stemmer and synonym list for every language
- compute-intensive (alignment, stemming, synonym lookup)

Topics

Overview: Manual Evaluation

manual eval

absolute

rate adequacy and fluency

relative

rank systems

→ comparisons

Overview: Automatic Evaluation

automatic

- Precision / Recall / F-Measure
  - ignores word order
- WER / TER
  - too strict
- BLEU
  - #1 metric!
  - all words equally important
- METEOR
  - language-dependent

Literature I

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan, Seiten 65–72.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. Sofia, Bulgaria, Seiten 1–44.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy, Seiten 249–256.

Literature II

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation (WMT)*. New York, NY, USA, Seiten 102–121.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*. Philadelphia, PA, USA, Seiten 311–318.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*.