Machine Translation Evaluation

Multilinguisme — lecture n°1

Guillaume Wisniewski
`guillaume.wisniewski@linguist.univ-paris-diderot.fr`

November 2019

Université de Paris & LLF

Context

Why Evaluating (Machine) Translation Quality?

Practitioners and Users

- how to compare alternative systems, choices for building and customizing off-the-shelves systems?
- quantify the effectiveness of using MT

Researchers

- evaluate models
- train our systems!

Translation Quality versus MT Quality

- Are human translators making mistakes?
  ↪ no, if you believe them
- Quality of human translation generally measured by number of words edited/corrected in the editing or proof-reading stages
- One of the main problem of translation studies!
- Can the same metrics be applied to MT systems?
  ↪ no, their output is not good enough ⊕ very different error profile

Usage scenario

Offline 'benchmark' testing of MT engine performance

- sample representative test documents with reference human translation available
- assess the performance on this particular dataset

Operational quality assessment at runtime

- MT engine is translating new source material
- Is the output 'good enough' for the underlying application?
- reference-less evaluation / confidence estimation

Common usage of reference-based evaluation

On the same dataset

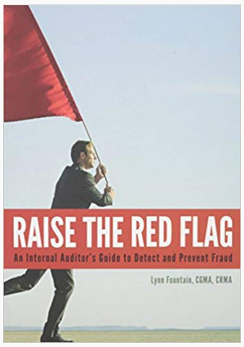
- compare two MT engines
- compare two versions of the same engine
  ↪ before and after customizing the engine
  ↪ before and after incremental development of the engine

On different datasets

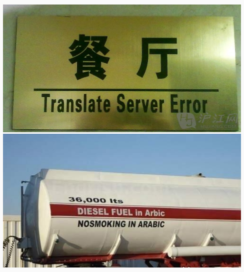
- compare MT engine performance
  ↪ across domains or types of input data
  ↪ on different sentence types, linguistic structures

Common usage of confidence evaluation

- identify / flag poorly translated segment during MT engine operation
  - ↪ can the translation be used as-it?
  - ↪ is it worth correcting the translation or is it better to translate it from scratch?


RAISE THE RED FLAG
An Internal Auditor's Guide to Detect and Prevent Fraud
Lynn Fountain, CGMA, CRMA

Be careful



User expectations on MT depends on their knowledge of:

- the source language
- the target language

Be careful

When you are using Google Translate to translate French into English your knowledge of English is good enough to decide whether you want to keep the translation or modify it. What about translating into Greek?

User expectations on MT depends on their knowledge of:

- the source language
- the target language

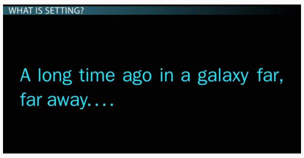
The difficulties of MT Evaluation

Major Issues

- Language variability: there is no single correct translation
- human evaluation is subjective
- how good is ‘good enough’?
- evaluation depends on the target application and context

Automatic Metrics

The setting


WHAT IS SETTING?

A long time ago in a galaxy far, far away. . . .

Given

- a reference translation
- a translation hypothesis

predicted by a MT system

How similar are the sentences / documents?

↪ fast and cheap, minimal human labor

↪ difficult to distinguish subtle differences between an hypothesis and a reference

Historical Metrics

Main ideas:

- same words in the hypothesis and in the reference(s)?
- in the same order?
- avoid adding extra words

⇒ use only surface information

↪ RI-style evaluation (Bleu, NIST, Meteor, MMS, ...)

↪ recall/precision on words

The Bleu score: intuition

References

❶ It is a guide to action that ensures that the military will forever heed Party commands.

❷ It is the guiding principle which guarantees the military forces always being under the command of the Party.

❸ It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The Bleu score: intuition

References

❶ It is a guide to action that ensures that the military will forever heed Party commands.

❷ It is the guiding principle which guarantees the military forces always being under the command of the Party.

❸ It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The Bleu score: intuition

References

❶ It is a guide to action that ensures that the military will forever heed Party commands.

❷ It is the guiding principle which guarantees the military forces always being under the command of the Party.

❸ It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The Bleu score: intuition

At the end:

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

↪ more common, longer *n*-grams in the second hypothesis ⇒ better translation

Formally i

- Brevity Penalty :

$$\mathrm{BP} = \begin{cases} 1 & \text{if hypothesis longer than reference} \\ e^{1 - \frac{\#r}{\#h}} & \text{otherwise} \end{cases} \tag{1}$$

↪ penalizes hypothesis that are 'too' long ≃ recall

- Modified n-gram precision :

$$p_n = \sum_{n\text{-gram} \in \text{hyp}} \frac{\#_{\text{clip}} n\text{-gram}}{\# n\text{-gram}} \tag{2}$$

with $\#_{\text{clip}}(a) = \min \{\#a \in h, \#a \in r\}$

↪ no reward for over-generated words

Formally ii

- Final score :

$$\textsc{Bleu} = \mathrm{BP} \times \exp \left(\sum_{n=1}^{N} w_n \log p_n \right)$$

where $\sum_{i=1}^{N} w_i = 1$

- usually $N = 4$ and $W_i = \frac{1}{4}$

More precisely: n-gram generator

```
def ngrams(sequence, n):
    sequence = iter(sequence)

    history = []

    while n > 1:
        history.append(next(sequence))
        n -= 1

    for item in sequence:
        history.append(item)
        yield tuple(history)
        del history[0]
```

More precisely: clipped *n*-gram precision

```
bag_ref = {(ngram, len(list(g)))
           for ngram, g in groupby(sorted(ngrams(ref, n)))}
bag_hyp = {(ngram, len(list(g)))
           for ngram, g in groupby(sorted(ngrams(hyp, n)))}

common = sum((min(bag_hyp[ngram], bag_ref.get(ngram, 0))
              for ngram in bag_hyp))
total = sum(bag_hyp.values())

return common / total
```

An historical note

- Introduced in (Papineni et al. 2002)
- First paragraph of the article:

  "[...] developers of machine translation systems need to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas. [...] We propose such an evaluation method in this paper."

  ↪ not an evaluation of translation quality or a way to compare systems!

  ↪ only designed to compare a system and its modification!
- Why using BP, exp, modified precision, ...

  ↪ they have tested several way to combine 'basic' information...

  ↪ ... and kept the one that best correlates with human judgments.

  But: no precise description of what has been tested ⊕ correlation measured on a small corpus that is not freely available.

The problem with BLEU (1)

Practical issues

- can only be computed at the corpus level

  ↪ a lot of application requires a score at the sentence level
- very hard to optimize: not decomposable, not differentiable

Questioning BLEU definition

- for an average hypothesis there are millions of possible variants (generated either by permuting or replacing n-grams) with a similar BLEU score ⇒ but they are not all grammatically or semantically plausible
- see e.g. (Callison-Burch, Osborne, and Koehn 2006) : *'We show that an improved Bleu score is neither necessary nor sufficient for achieving an actual improvement in translation quality.'*

The problem with BLEU (2)

From an end-user point of view

- BLEU scores are not fully comparable across languages or even across different benchmarks for the same language
- not easily interpretable by most translation professionals
- scores depends on the implementation
  - ↪ impact of tokenization
  - ↪ implementation details (several ad-hoc decisions)

  ⇒ *de facto* standard: (Post 2018)

Score interpretation

- score over 30 generally reflect understandable translations
- scores over 50 generally reflect good and fluent translations

⇒ most papers report improvement $\simeq 1$ BLEU point

At the end...

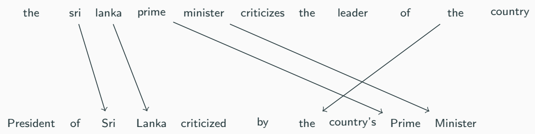
Like it or not, you have to use it (Blatz et al. 2004)

↪ there is not a single MT paper that does not report BLEU scores

Meteor

- Metric for Evaluation of Translation with Explicit Ordering [Lavie and Denkowski, 2009]
- rely on unigram recall and precision
  - ↪ align/match words of the hypothesis and of the reference
- matching takes into account translation variability via word inflection variations, synonymy and paraphrasing matches
- direct penalty for word order: how fragmented is the matching?
- weighted metrics: the weights of the difference components can be optimized to improve correlation with human judgments

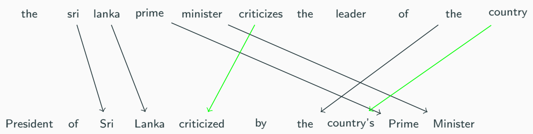
Alignment



the sri lanka prime minister criticizes the leader of the country

President of Sri Lanka criticized by the country's Prime Minister

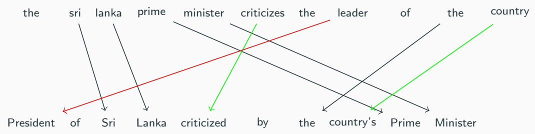
Alignment



the sri lanka prime minister criticizes the leader of the country

President of Sri Lanka criticized by the country's Prime Minister

Alignment



the sri lanka prime minister criticizes the leader of the country

President of Sri Lanka criticized by the country's Prime Minister

↪ different kind of matches, with different weights

↪ optimal search is NP-complete (but clever search with pruning is very fast and has near optimal results)

The Full Meteor Metric

Fragmentation

- to take fluency into account
- $\text{frag} = \frac{\#\text{`group' of word that are in matches}-1}{\#\text{words in matches}-1}$

Final Score

- discounting factor: $\text{DF} = \gamma \times \text{frag}^{\beta}$
- F_α score: $F_\alpha = \frac{P \times R}{\alpha \cdot P + (1-\alpha) \cdot R}$
- original parameter settings: $\alpha = 0.9, \beta = 3.0, \gamma = 0.5$
- final score: $F_\alpha \cdot (1 - \text{DF})$

Meteor example

Example

- **Reference:** “the Iraqi weapons are to be handed over to the army within two weeks”
- **Hypothesis:** “in two weeks Iraq’s weapons will give army”

Meteor example

Example

- **Reference:** “the Iraqi weapons are to be handed over to the army within two weeks”
- **Hypothesis:** “in two weeks Iraq’s weapons will give army”

↪ Precision =

Meteor example

Example

- **Reference:** “the Iraqi weapons are to be handed over to the army within two weeks”
- **Hypothesis:** “in two weeks Iraq’s weapons will give army”

↪ Precision = $\frac{5}{8} = 0.625$

Meteor example

Example

- **Reference:** “the Iraqi weapons are to be handed over to the army within two weeks”
- **Hypothesis:** “in two weeks Iraq’s weapons will give army”

↪ Precision = $\frac{5}{8} = 0.625$

↪ Recall =

Meteor example

Example

- **Reference:** “the Iraqi weapons are to be handed over to the army within two weeks”
- **Hypothesis:** “in two weeks Iraq’s weapons will give army”

↪ Precision = $\frac{5}{8} = 0.625$

↪ Recall = $\frac{5}{14} = 0.357$

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision $= \frac{5}{8} = 0.625$

↪ Recall $= \frac{5}{14} = 0.357$

↪ Fragmentation =

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision $= \frac{5}{8} = 0.625$

↪ Recall $= \frac{5}{14} = 0.357$

↪ Fragmentation $= \frac{3-1}{5-1} = 0.5$

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision $= \frac{5}{8} = 0.625$

↪ Recall $= \frac{5}{14} = 0.357$

↪ Fragmentation $= \frac{3-1}{5-1} = 0.5$

Weighted combination: 0.3498

The problem with Meteor

- easier to interpret?
- rely on external resources (e.g. paraphrase table) that are not always available
- computational cost
- fuzzy matches have very low impact

TER

- Translation Edit (Error) Rate (Snover et al. 2009)
- edit-based measure: number of edits to transform hypothesis into reference
- add the notion of 'block movements' as single edit operation

  but NP-complet
- exact matches only — extension TERp: near-matches ⊕ weights
- rough estimate of post-editing effort

Meta-Evaluation: how good are MT metrics

Comparing Metrics

How do we know if a metric is better?

1. Better correlation with human judgments of MT output
2. Reduced score variability on MT outputs that are ranked equivalent by humans
3. Higher and less variable scores on scoring human translations against the reference translations

↪ several challenges to find the best metric

↪ still a hot topic in MT research

Manual Evaluation

Why Perform Human Evaluation?

- automatic MT metric are not sufficient:
  - ↪ not interpretable
  - ↪ biased
  - ↪ no possibility of error analysis
- need for reliable human measure to evaluate automatic metric

Difficulties

- time & cost
- reliability and consistency: difficulty in obtaining high-levels of intra and inter-coder agreement

Historical Human Metrics

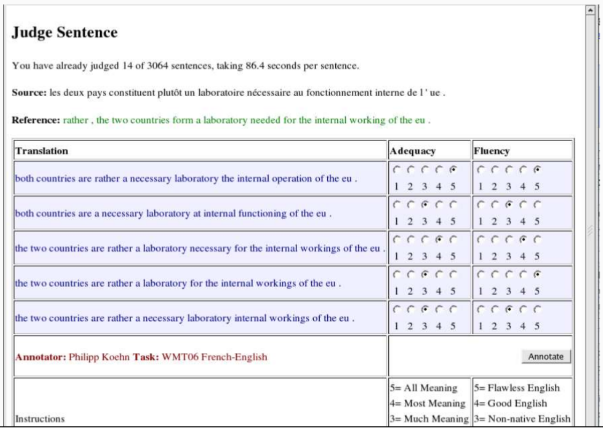
- Adequacy: is the meaning translated correctly?
  - ↪ By comparing MT translation to a reference translation (or to the source)?
- Fluency: is the output grammatical and fluent?
  - ↪ By comparing MT translation to a reference translation, to the source, or in isolation?
- same scale: $[\![1, 5]\!]$
- initiated during DARPA MT evaluation in the mid-1990s
- main issues: definition of scales, agreement, normalization across judges

Fluency and Adequacy Scales

Adequacy	
1	no meaning
2	little meaning
3	much meaning
4	most meaning
5	all meaning

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

With a picture:



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	5= Flawless English 4= Good English 3= Non-native English

Let's try it!

- **Source:** N'y aurait-il pas comme une vague hypocrisie de votre part?
- **Reference:** Is there not an element of hypocrisy on your part?
- **System 1:** Would it not as a wave of hypocrisy on your part?
- **System 2:** Is there would be no hypocrisy like a wave of your hand?
- **System 3:** Is there not as a wave of hypocrisy from you?

Consistency of judgments

Inter-rater agreement	
	κ
2011	0.40
2012	0.33
2013	0.26
2014	0.37

Intra-rater agreement	
	κ
2011	0.58
2012	0.41
2013	0.48
2014	0.52

⇒ very low in both cases

A new proposition for human evaluation (Graham, Baldwin, and Mathur 2015)

Modus operandi

1. each sentence is rated by n humans on a [0, 100] scale (monolingual evaluation)
2. normalize the scores of each raters (i.e. consider the z-score)
3. define the quality of the translation by its mean rating

Why?

- law of large numbers ⇒ the mean to the expected value
- axiomatic choice: the expected value is the true translation quality.

The unreasonable effectiveness of data

- only one difference with the historical human metrics: collect as many ratings as possible (in practice $\simeq 15$)
- no need to define the notion of 'translation quality' formally ⇒ let it emerge from the data.

At the end

Method that is used today in most evaluation campaign

References

- Blatz, John et al. (2004). “Confidence Estimation for Machine Translation”. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING, pp. 315–321. URL: https://www.aclweb.org/anthology/C04-1046.
- Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). “Re-evaluating the Role of Bleu in Machine Translation Research”. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E06-1032.

- Graham, Yvette, Timothy Baldwin, and Nitika Mathur (2015). “Accurate Evaluation of Segment-level Machine Translation Metrics”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, pp. 1183–1191. DOI: 10.3115/v1/N15-1124. URL: https://www.aclweb.org/anthology/N15-1124.
- Papineni, Kishore et al. (2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.

- Post, Matt (2018). “A Call for Clarity in Reporting BLEU Scores”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.
- Snover, Matthew et al. (2009). “Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics, pp. 259–268. URL: https://www.aclweb.org/anthology/W09-0441.