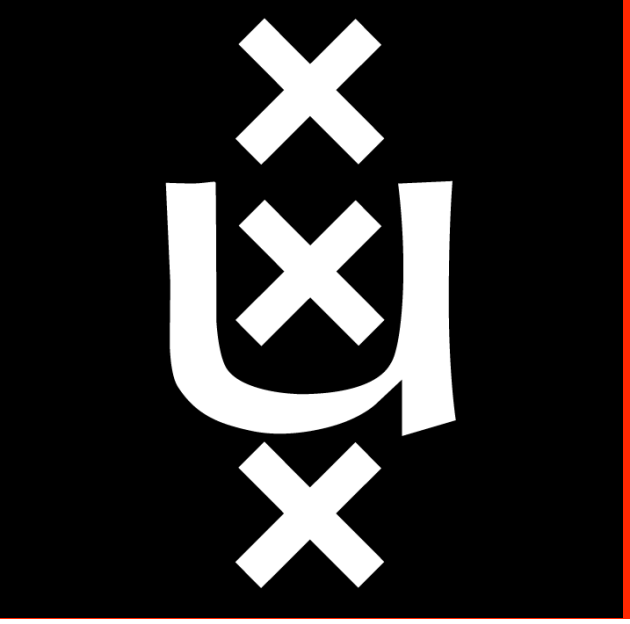


Christof Monz

Applied Language Technology
Evaluation of Machine Translation

Today's Class

- Evaluation of machine translation
- Automated evaluation metrics

Evaluation of Machine Translation

- Thorough evaluation of MT is important to answer a number of questions
  - How good is our system or approach?
  - How do two approaches compare?
  - How much did our system improve?
  - How can optimize our system?
  - Which is the best approach for a given task?
- We would like the answers to these questions to be
  - reliable
  - quantifiable
  - repeatable

Types of Evaluation

- Intrinsic evaluation: How good is the MT system by itself (in isolation)?
  - Quantitative evaluation
    - How good is the translation quality of the overall system?
  - Qualitative evaluation
    - What mistakes does the system make? E.g., agreement errors, dropping verbs, OOV rates, ... ($\rightarrow$ error analysis)
- Extrinsic evaluation: How much does the MT system help to perform a task?
  - Involves end-users carrying out a task that requires MT in the pipeline
  - E.g., reading comprehension, being able to answer questions about a document
  - Also includes usefulness for further data mining, e.g., named entity recognition

Types of Evaluation

- Human evaluation
  - Requires human assessors that are native speakers of the target language and fluent in the source language (just like translators)
  - Measures ultimate goal
  - Time consuming
  - Not-reusable as it depends on the specific translations
- Automated evaluation
  - Requires an automated metric that correlates with human judgments
  - Cheap and fast
  - Reusable
  - Reliability can be an issue

Human Evaluation of MT

- Typically measured along two dimensions:
  - Adequacy: How much of the meaning of the original sentence is preserved?
  - Fluency: How fluent (and grammatical) is the translation?
- Measured along Likert 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

Evaluation Exercise

باكو - رويترز: قال مسؤول أميركي أمس ان الولايات المتحدة تجر حادثات مع العراق وتركيا لجمع استثمارات بهدف استئناف انتاج الغاز العراقي وتصدي ره الى أوروبا.
Human: Baku – Reuters: An American official said yesterday that the United States is holding talks with Iraq and Turkey to collect investments aimed at resuming the production of Iraqi gas and exporting it to Europe.
MT: Baku-Reuters: An American official said yesterday that the United States is holding talks with Iraq and Turkey to collect investments with a view to resuming the production of Iraqi gas and exporting it to Europe.

Evaluation Exercise

باكو - رويترز: قال مسؤول أميركي أمس ان الولايات المتحدة تجر حادثات مع العراق وتركيا لجمع استثمارات بهدف استئناف انتاج الغاز العراقي وتصدي ره الى أوروبا.

Human: Baku – Reuters: An American official said yesterday that the United States is holding talks with Iraq and Turkey to collect investments aimed at resuming the production of Iraqi gas and exporting it to Europe.

MT: Baku-Reuters: An American official said yesterday that the United States is holding talks with Iraq and Turkey to collect investments with a view to resuming the production of Iraqi gas and exporting it to Europe.

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

Evaluation Exercise

هو في حاجة إلى بطانة صادقة تساعده لا إلى وحوش كاسرة أنشبت خابها ف الكراسي وتطمع في المزيد
Human: He needs an honest entourage that would help him, not vicious monsters that put their claws on the chairs and are greedy for more.
MT: It is in need of sincere entourage does not help to monsters in a family claws chairs and seeks to win more.

Evaluation Exercise

هو في حاجة إلى بطانة صادقة تساعده لا إلى وحوش كاسرة أنشبت خابها ف الكراسي وتطمع في المزيد
Human: He needs an honest entourage that would help him, not vicious monsters that put their claws on the chairs and are greedy for more.
MT: It is in need of sincere entourage does not help to monsters in a family claws chairs and seeks to win more.

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

Evaluation Exercise

أستغربُ صمت الإعلام العربي على الزيارة غير المرحب بها، التيامها إعصار غونو لسواحل الخليج.
Human: I feel strange about the silence of the Arab media regarding the unwelcomed visit of Hurricane Gonu to the Gulf coasts.
MT: At the silence of the Arab media on the visit, not amicability paid by hurricane for the Gulf coast.

Evaluation Exercise

أستغربُ صمت الإعلام العربي على الزيارة غير المرحب بها، التيامها إعصار غونو لسواحل الخليج.
Human: I feel strange about the silence of the Arab media regarding the unwelcomed visit of Hurricane Gonu to the Gulf coasts.
MT: At the silence of the Arab media on the visit, not amicability paid by hurricane for the Gulf coast.

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

Human Evaluation

- Absolute judgments are somewhat unreliable
- Different users have different 'standards'
  - User scores can be normalized by 'strictness' of evaluator
- Make task easier for judges
  - Rank translations instead of absolute judgments
  - Rank segments instead of entire sentences
- Measure effort instead of qualitative judgments
  - How much effort is needed to correct a translation?

Human Evaluation: Sentence Ranking

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst

Human Evaluation: Segment Ranking

Rank Segments

You have judged 25 sentences for **WMT07 German-English News Corpus**, 190 sentences total taking 64.9 seconds per sentence.

Source: Können die USA **ihre Besetzung aufrechterhalten**, wenn sie dem irakischen Volk nicht Nahrung, Gesundheitsfürsorge und andere grundlegende Dienstleistungen anbieten können?

Reference: Can the US **sustain its occupation** if it cannot provide food, health care, and other basic services to Iraq's people?

Translation	Rank
The United States can **maintain its employment** when it the Iraqi people not food, health care and other basic services on offer?.	○ 1 Worst ◉ 2 ○ 3 ○ 4 ○ 5 Best
The US can **maintain its occupation**, if they cannot offer the Iraqi people food, health care and other basic services?	○ 1 Worst ○ 2 ○ 3 ○ 4 ◉ 5 Best
Can the US **their occupation sustained** if it to the Iraqi people not food, health care and other basic services can offer?	○ 1 Worst ○ 2 ◉ 3 ○ 4 ○ 5 Best
Can the United States **maintain their occupation**, if the Iraqi people do not food, health care and other basic services can offer?	○ 1 Worst ○ 2 ○ 3 ◉ 4 ○ 5 Best
The United States **is maintained**, if the Iraqi people, not food, health care and other basic services can offer?	○ 1 Worst ◉ 2 ○ 3 ○ 4 ○ 5 Best
Annotator: ccb **Task:** WMT07 German-English News Corpus	
Instructions: Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the highlighted part** of each translation. *Please note that segments are selected automatically, and they should be taken as an approximate guide. They might include extra words on either end that are not in the actual alignment, or miss words.*	

Human Evaluation: Editing Effort

Your edited translations

The machine translations

The shortage of snow in mountain worries the hoteliers

The shortage of snow in mountain worries the hoteliers

Edited No corrections needed Unable to correct Reset

The deserted tracks are not putting down problem only at the exploitants of skilift.

The deserted tracks are not putting down problem only at the exploitants of skilift.

Edited No corrections needed Unable to correct Reset

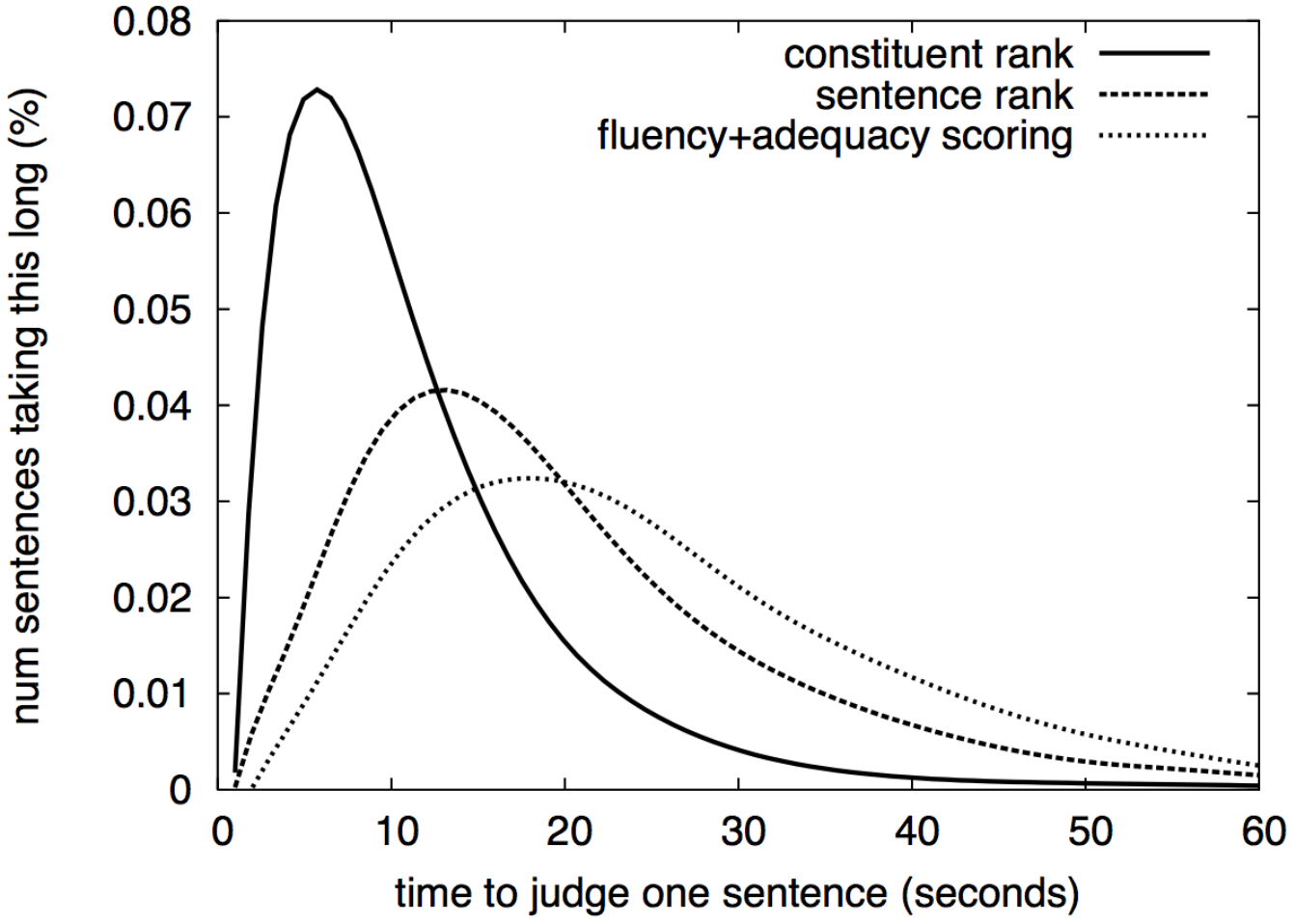
The lack of snow deters the people to reserving their stays at the ski in the hotels and pension.

The lack of snow deters the people to reserving their stays at the ski in the hotels and pension.

Edited No corrections needed Unable to correct Reset

Time Spent on Human Evaluation

- Human evaluation is a time-intensive task



Criteria for Automated Evaluation Metrics

- Reusable
  - If we evaluation system A at two different points in time, we get the same result
- Fast
  - Evaluation is not too time-consuming allowing for rapid development and iterative parameter tuning
- Non-parameterized
  - Some metrics have weighted parameters: results can differ by using different weights (see also re-usability)
- High correlation with human judgments
  - A ranking of systems created by human judges should be reflected by the ranking of an automated metric
  - Use Kendall's Tau, Spearman correlation, etc.
- Robust
  - If system A ranks higher than system B on test set X, A also ranks higher B on related but different test set Y

Criteria for Automated Evaluation Metrics

- Intuitive
  - Ideally the metric measures something that is intuitive
  - The score should have an intuitive interpretation

Automated MT Evaluation Metrics

- Word Error Rate (WER)
  - Similar to string edit distance, but using words instead of characters
  - Counts `insert`, `delete`, and `replace` operations to turn automated translation into human translation
- Translation Error Rate (TER)
  - Similar to WER but allows for continuous segments to be treated as one edit operation
- METEOR
  - Precision and recall metric, allowing for flexible matching (stems and synonyms)
  - Language-dependent
- BLEU
  - Precision-oriented metric counting overlaps between translation and human reference

Evaluation Data

- All automated evaluation metrics require evaluation (test) data, aka 'benchmark' or 'ground truth'
- A test set consists of
  - A fixed set of sentences in the source language (sentence boundaries are fixed)
  - A set of human translation for each source sentence
    - Also know as reference translations or simply references
    - Ideally several translations of the same source sentence by different translators
  - Selection of data should
    - cover several domains: sports, politics, etc.
    - cover several styles: formal news, informal user-generated content
  - Data should be of sufficient size to cover different phenomena, domains, styles, etc.
    - Rule of thumb: 1-2K for multiple references, 2K for 1 reference

BLEU

- Bilingual Evaluation Understudy (BLEU), introduced by Roukos and Papineni (2003)
- BLEU is a precision-oriented metric
  - How many of the n-grams occurring in the translation occur in any of the reference translations?
  - Considers n-grams of several lengths, typically 1 to 4-grams
- N-gram precision: $p(i) = \frac{\text{correct}_i}{\text{total}_i}$
  - For each n-gram occurrence $\bar{w}_i$ of order i in the translation,
    - $\text{correct}_i{+}{+}$ and $\text{total}_i{+}{+}$: if $\bar{w}_i$ occurs in any of the reference translations
    - $\text{total}_i{+}{+}$: else
- Brevity penalty (BP): Compensates for the tendency of shorter translations having higher precisions

$$\text{BP}(l_t, l_r) = \begin{cases} 1 & \text{if } l_t \geq l_r \quad (l_r = \text{reference length}) \\ \exp(1 - \frac{l_r}{l_t}) & \text{if } l_t < l_r \quad (l_t = \text{translation length}) \end{cases}$$

BLEU Example

src:	die katze lag auf der matte .
ref-1:	the cat lied on the mat .
ref-2:	the cat sat on the mat .
ref-3:	the cat was lying on the mat .
ref-4:	the cat lied on this mat .

- Candidate translation t: the cat sat on a mat .

  $p(1) = \frac{6}{7}$: the; cat; sat; on; a; mat; .

  $p(2) = \frac{4}{6}$: the cat; cat sat; sat on; on a; a mat; mat .

  $p(3) = \frac{3}{5}$: the cat sat; cat sat on; sat on a; on a mat; a mat .

  $p(4) = \frac{1}{4}$: the cat sat on; cat sat on a; sat on a mat; on a mat .

$$\text{BLEU}(t, R_f) = \text{BP}(7,7) \cdot \prod_{i=1}^{n} p(i)^{\frac{1}{n}} = \text{BP} \cdot \frac{6}{7}^{\frac{1}{4}} \cdot \frac{4}{6}^{\frac{1}{4}} \cdot \frac{3}{5}^{\frac{1}{4}} \cdot \frac{1}{4}^{\frac{1}{4}}$$

$$= \text{BP}(7,7) \cdot 0.5411 = 1 \cdot 0.5411 = 0.5411$$

BLEU: Adjustments

- N-gram precision: $p(i) = \frac{\text{correct}_i}{\text{total}_i}$
- Candidate translation t: the the the the the the the

  $$p(1) = \frac{\text{correct}_1}{\text{total}_1} = \frac{7}{7} = 1$$
- Use clipped-counts: Let $\text{ref\_count}(\bar{w}_i)$ be the maximum number of times $\bar{w}_i$ occurs in any individual reference and $\text{trans\_count}(\bar{w}_i)$ the number of times $\bar{w}_i$ occurs in the translation
- For each n-gram $\bar{w}_i$ of order i in the translation,

  $$\text{correct\_clipped}_i += \min(\text{trans\_count}(\bar{w}_i), \text{ref\_count}(\bar{w}_i))$$
- $p(i) = \frac{\text{correct\_clipped}_i}{\text{total}_i}$

BLEU: Adjustments

- The brevity penalty compares the length of the translation candidate with the length of the reference translation
- If we have multiple reference translations, there are multiple ways to define l_r:
  - Shortest reference
  - Average reference length
  - Closest reference: The length of the reference which is closest in length to the translation

BLEU: Granularity

- BLEU scores can be computed on the
  - sentence-level
  - document-level
  - corpus-level
- Sentence-level BLEU is rather unstable, translation candidates with minor differences can receive very different sentence-level BLEU scores
- Better apply BLEU on the document or corpus level
- BLEU formulation remains unchanged, but all counts and lengths are computed over the entire document or corpus

Comments on BLEU

- BLEU generally correlates relatively well with human judgments
- By far the most commonly used MT evaluation metric
- Absolute BLEU scores in isolation are not very meaningful
- Comparison of BLEU scores across language pairs not meaningful
- BLEU is less useful for translating into morphologically rich languages
- Good translations that are dissimilar to all reference translations are penalized

Recap

- Evaluation of machine translation
  - Types of evaluation
  - Human evaluation
- Automated evaluation metrics
  - criteria
  - test data
  - BLEU