CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Natural Language Processing

Week8: Machine Translation Evaluation

Dr. Haithem Afli

Haithem. afli@cit.ie

@AfliHaithem

2019/2020

Recap and Quiz

- What are the three main components of an SMT system?

- What is the difference between fluency and adequacy?

Content

1. **Introduction**
2. Human Evaluation
3. Automatic Evaluation
4. Task-based Evaluation

MT Evaluation

- How good is a given MT system?
- Why is this a difficult question to answer?

- 这个 机场 的 安全 工作 由 以色列 方面 负责 .
- 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.
  - (From 2001 NIST Evaluation)

- 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.
  - (From 2001 NIST evaluation)

Goals for MT Evaluation

- Meaningful: score should give intuitive interpretation of translation quality
- Consistent: repeated use of metric should lead to same results
- Correct: metric must rank better systems higher
- Low cost: reduce time and money spent to carry out evaluation
- Tunable: automatically optimise system performance towards metric

Other Evaluation Criteria

Other issues besides translation quality

- Speed: is the system fast enough in practice?
- Size: fits into memory of available machines (e.g., handheld devices)
- Integration: into existing workflows
- Customisation: can be adapted to user's needs

Different Types of Evaluation

1. Subjective judgments by human evaluators
2. Automatic evaluation metrics
3. Task-based evaluation, e.g.
   - how much post-editing effort?
   - does the information come across?

Content

1. Introduction
2. **Human Evaluation**
3. Automatic Evaluation
4. Task-based Evaluation

Human Evaluation

- Given
  - MT output
  - source and/or reference translation
    - **Reference translation**: a translation produced by a trained translator (human)
- Task: **assess the quality of MT output**

Human Evaluation

English-to-Irish example

- MT Output: *Tá mé múinteoir*
- Source: *I am a teacher*
- Reference Translation: *Tá mé i mo mhúinteoir*
- Task: **assess the quality of the MT output given the source and reference translation**

Human Evaluation

- 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 a good fluent sentence in the target language?
  - This involves both grammatical correctness and idiomatic word choices

Human Evaluation

Evaluating Adequacy

5 all meaning

4 most meaning

3 much meaning

2 little meaning

1 none

Human Evaluation

Evaluating Fluency

5 flawless English

4 good English

3 non-native English

2 disfluent English

1 incomprehensible

Annotation Tool for Human Evaluation

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

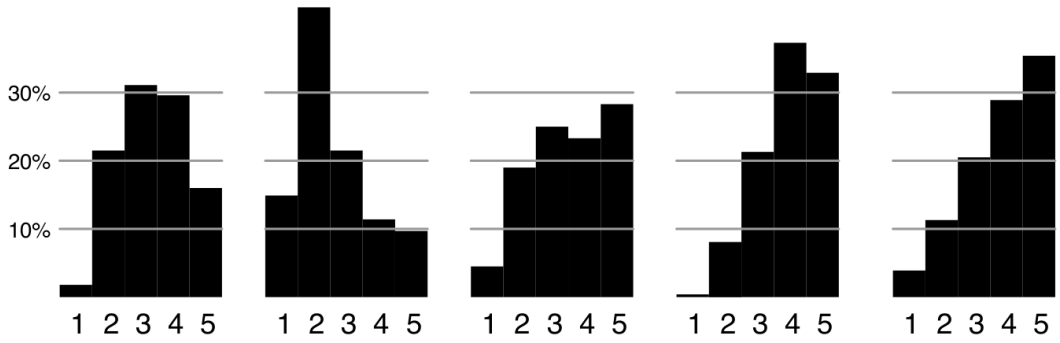
Human Evaluation

- Some evaluators more lenient than others, normalise average, given judgements

Human Evaluation

- Some evaluators more lenient than others, normalise average, given judgements

- Still, evaluators disagree (distributions)



(WMT 2006 evaluation task)

Human Evaluation

- Another evaluation task carried out by judges is to rank different translations
- Judges are more likely to agree with each other on this task
  - Relative vs absolute judgement
  - Inter-annotator agreement

Ranking by Pairwise Comparison

- Instead of giving a score to each system,

we try to rank all the candidate systems according to their translation quality

- Human evaluators asked to do pair-wise

comparisons between MT outputs for each sentence

- The full ranking is generated based on the

pair-wised comparison results

Afganistanci su platili cijenu opskurantizma tih seljaka organizacijom Al-Kaide, no njihova situacija se do danas nije poboljšala. **Bivši Mujahidin, afganistanska vlada i trenutni Talibani su se sjedinili u želji da održe žene u podređenom položaju.** Glavni anti-sovjetski ratni vođe vratili su se na vlast 2001.

— Source

Afghanis paid the price of the obscurantism of these peasants by the organisation of Al-Qaeda, but their situation has not improved today. **Former Mujahidin, the Afghan Government and the current Taliban are allied in the desire to keep women in an inferior position.** The main anti-Soviet war leaders returned to power in 2001.

— Reference

○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5

Former Mujahidin, Afghan government and the Taliban have joined themselves in order to keep women in a subordinate position.

— Translation 1

○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5

A former Mujahidin, Afghan Government and the current Taliban are joined in the desire to keep women in a subordinate position.

— Translation 2

○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5

A former Mujahidin, the Afghan government and the current Taliban are united in the desire to keep women in a subordinate position.

— Translation 3

○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5

Former Mujahidin, the Afghan government and the Taliban were to unite in the desire to provide women in a subordinate position.

— Translation 4

○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5

Former Mujahidin, Afghan government and the Taliban are to be merged in order to keep women in a subordinate position.

— Translation 5

Submit Reset Flag Error

(Appraise evaluation toolkit, annotation)

Problems

- How can we get the overall ranking given all pair-wise comparisons?
- How can we get the overall ranking given part of pair-wise comparisons?
- How can we get the overall ranking if we ask the human evaluators to rank 3-5 MT results each time (if there are more than 3-5 MT systems)?

»abu_m3_exp5_hren« Status Overview

User name	Overall completion	Average duration
[redacted]	100/1000	87.66 sec
Combined	100/1000	87.66 sec

Results Overview

1,1,1,1	1,1,1,2	1,1,2,1	1,1,2,3	1,1,3,2	1,2,1,1	1,2,1,2	1,2,2,2	1,2,3,1	1,2,3,2	1,2,3,4	1,2,4,3	1,3,1,2	1,3,2,1	1,3,2,2	1,3,2,4	1,3,4,2	1,4,3,2
9 (9.00%)	3 (3.00%)	2 (2.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	4 (4.00%)	2 (2.00%)	1 (1.00%)	2 (2.00%)	3 (3.00%)	3 (3.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%

(Appraise evaluation toolkit, status)

»abu_m3_exp5_hren« Status Overview

User name	Overall completion	Average duration
fklubicka	100/1000	87.66 sec
Combined	100/1000	87.66 sec

Results Overview

1,1,1,1	1,1,1,2	1,1,2,1	1,1,2,3	1,1,3,2	1,2,1,1	1,2,1,2	1,2,2,2	1,2,3,1	1,2,3,2	1,2,3,4	1,2,4,3	1,3,1,2	1,3,2,1	1,3,2,2	1,3,2,4	1,3,4,2	1,4,3
9 (9.00%)	3 (3.00%)	2 (2.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	4 (4.00%)	2 (2.00%)	1 (1.00%)	2 (2.00%)	3 (3.00%)	3 (3.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00%)	1 (1.00

(Appraise evaluation toolkit, status)

#	Score	Range	System
1	0.605	1-2	Google
2	0.540	1-4	Abu-MaTran (best combo)
3	0.519	1-4	Abu-MaTran (best individual)
4	0.506	2-4	Microsoft
5	0.330	5-5	Yandex

Table 10: Human ranking with Expected Wins for experiment 5, Croatian→English.

Content

1. Introduction
2. Human Evaluation
3. **Automatic Evaluation**
4. Task-based Evaluation

Automatic Evaluation Metrics

Computer program that computes the quality of translations

Advantages

- low cost
- tunable
- consistent (deterministic)

Automatic Evaluation Metrics

Basic strategy

- Input: MT output
- Input: human reference translation
- Output: a score which represents the similarity between the MT output and the human reference

Word Error Rate (WER)

Minimum number of editing operations to transform an MT output to a reference translation

- match: words match, no cost
- substitution: replace one word with another
- insertion: add word
- deletion: drop word

Word Error Rate (WER)

Levenshtein distance: minimum number of operations

$$\mathrm{WER} = \frac{insertions + deletions + substitutions}{reference\ length}$$

Word Error Rate (WER)

Reference translation:

Israeli officials are responsible for airport security

System output:

Israeli official responsible airport is security

WER score?

Word Error Rate (WER)

Reference translation:

Israeli officials are responsible for airport security

System output:

Israeli official responsible airport is security

Insertions	are, for	2
Deletions	is	1
substitutions	official →officials	1

WER score: 4/7

Discussion

WER Calculation

- Problem: Given a reference translation and an MT system output, how can we calculate the WER score?

WER Calculation

	Israeli	officials	are	responsible	for	airport	security
Israeli							
official							
responsible							
airport							
is							
security							

WER Calculation

	Israeli	officials	are	responsible	for	airport	security
Israeli							
official							
responsible							
airport							
is							
security							

WER Calculation

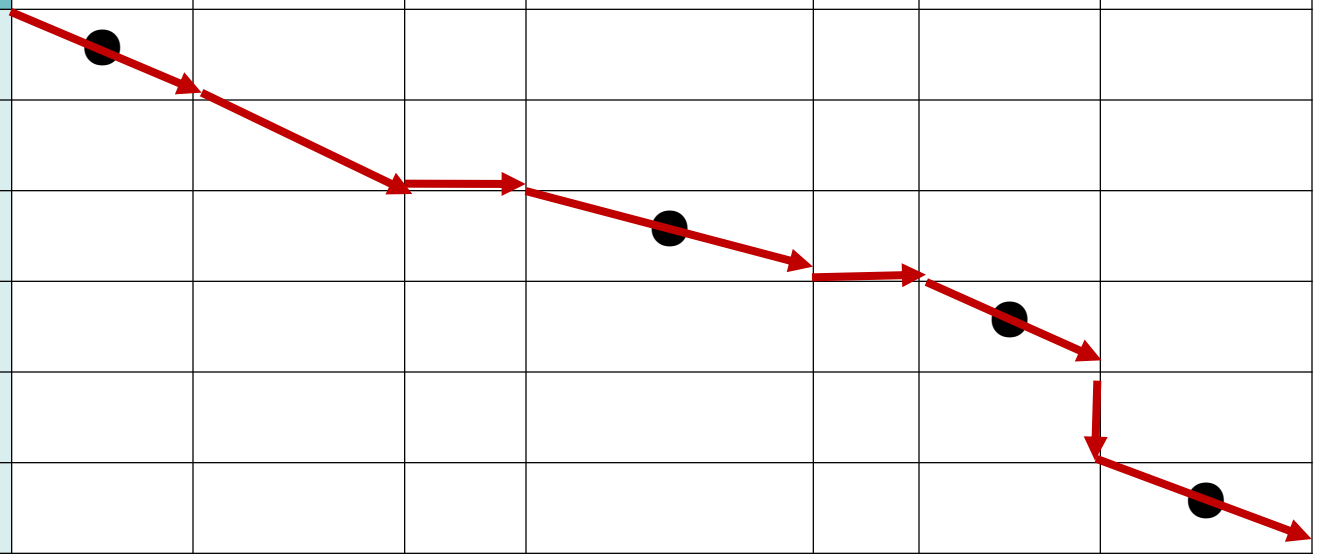
	Israeli	officials	are	responsible	for	airport	security
Israeli	●						
official							
responsible				●			
airport						●	
is							
security							●

WER Calculation

	Israeli	officials	are	responsible	for	airport	security
Israeli	●						
official							
responsible				●			
airport						●	
is							
security							●

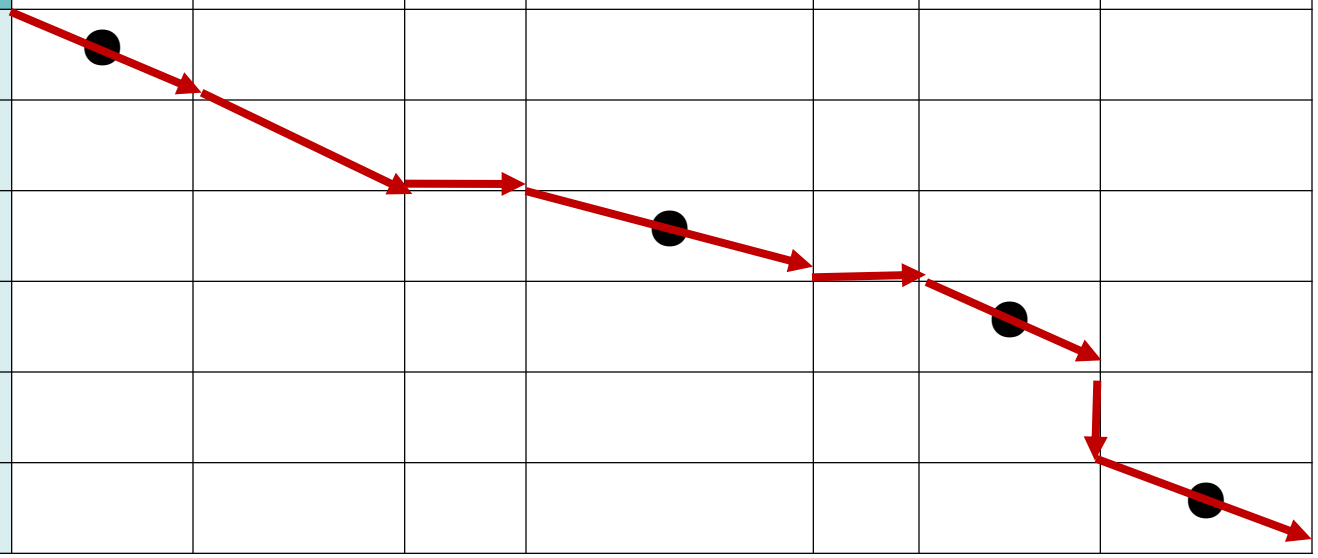
WER Calculation

	Israeli	officials	are	responsible	for	airport	security
Israeli							
official							
responsible							
airport							
is							
security							

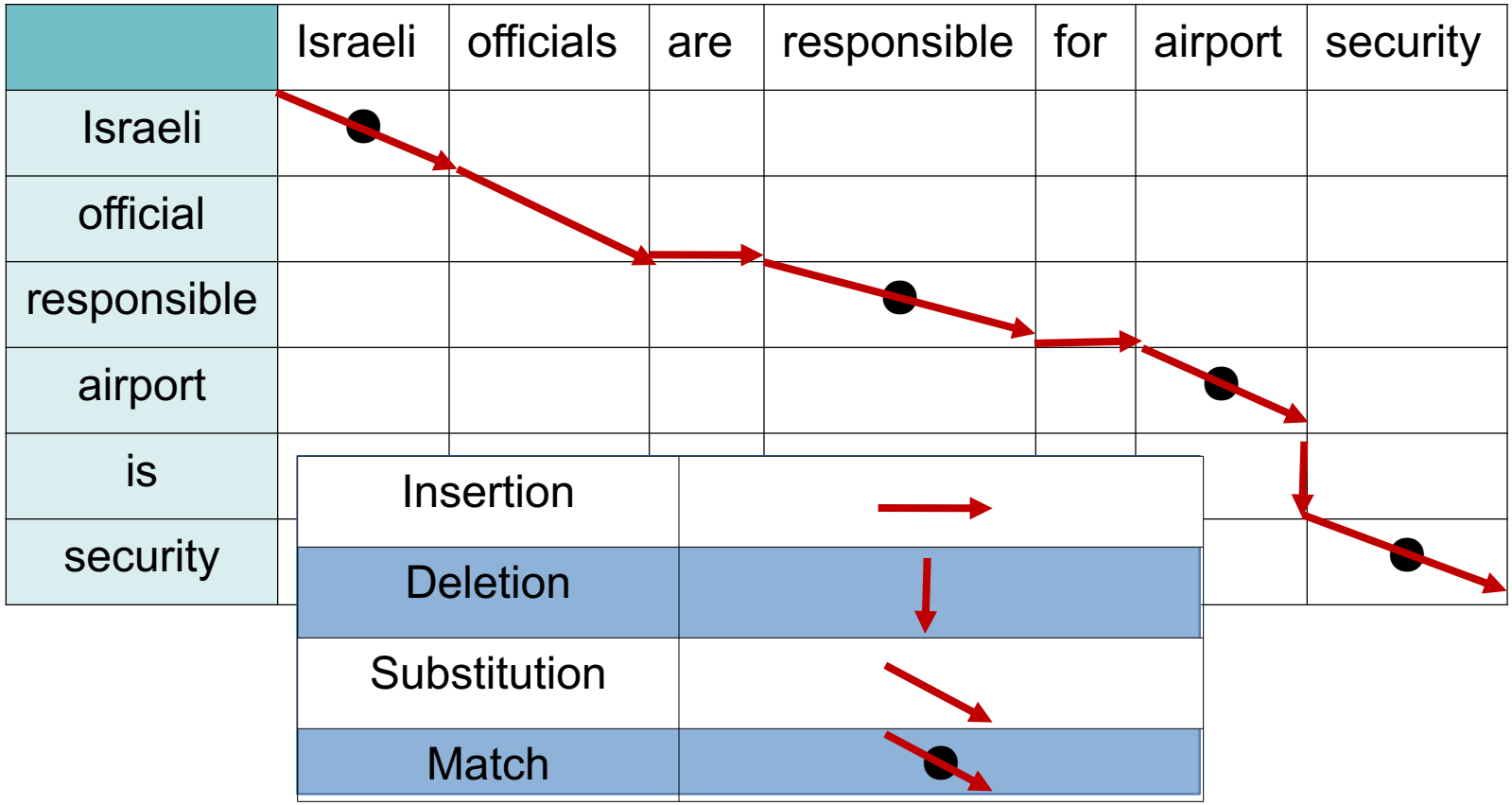


WER Calculation

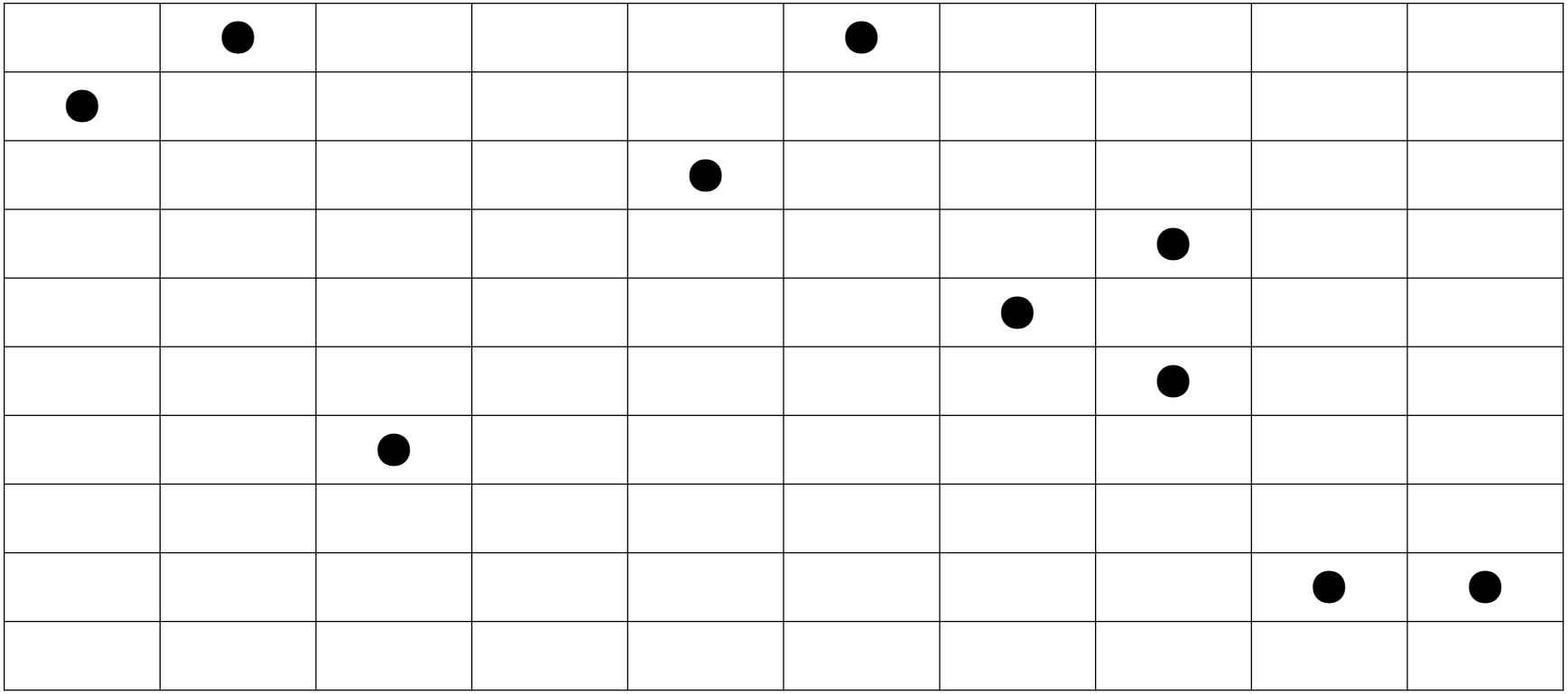
	Israeli	officials	are	responsible	for	airport	security
Israeli							
official							
responsible							
airport							
is							
security							



WER Calculation



Problem



Word Error Rate (WER)

Reference translation:

Israeli officials are responsible for airport security

System output:

This airport’s security is the responsibility of the Israeli security officials

Good translation but in opposite order to the reference translation -> **low WER score**

TER

Translation Error Rate is an error metric for machine translation that messures the number of edits required to change a system output into one of the references with additional costs for shifts of word sequences.

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

TER

```
REF: a b c d e f   c
HYP: a     d e   b c f
```

The words “b c” in the hypothesis can be shifted to the left to correspond to the words “b c” in the reference, because there is a mismatch in the current location of “b c” in the hypothesis, and there is a mismatch of “b c” in the reference.

Example from (Snover et al. 2006)

TER

```
REF: a b c d e f   c
HYP: a     d e   b c f
```

After the shift

the hypothesis is changed to:

```
REF: a b c d e f c
HYP: a b c d e f
```

BLEU

- N-gram overlap between MT output and reference translation
- Compute n-gram overlap for n = 1…4
- Typically computed over the entire corpus, not single sentences

N-grams

An n-gram is a sequence of words of order *n*

N-gram Example

"The cat sat on the mat"

? 1-grams (or unigrams)

? 2-grams (or bigrams)

? 3-grams (or trigrams)

? 4-grams

N-gram Example

"The cat sat on the mat"

6 1-grams (or unigrams)

- The, cat, sat, on, the, mat

5 2-grams (or bigrams)

4 3-grams (or trigrams)

3 4-grams

N-gram Example

"The cat sat on the mat"

5 1-grams (or unigrams)
- The, cat, sat, on, the, mat

5 2-grams (or bigrams)
- The cat, cat sat, sat on, on the, the mat

4 3-grams (or trigrams)
- The cat sat, cat sat on, sat on the, on the mat

3 4-grams
- The cat sat on, cat sat on the, sat on the mat

BLEU

$$\text{BLEU} = \min\left(1, \frac{output\ length}{reference\ length}\right) \left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

BLEU

$$\text{BLEU} = \underbrace{\min\left(1, \frac{output\ length}{reference\ length}\right)}_{\text{Brevity Penalty}} \underbrace{\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}}_{\text{N-gram Overlap}}$$

Multiple References

To account for variability, we can use multiple reference translations

- n-grams may match in any of the references
- closest reference length is used for brevity penalty

BLEU: Example

Example

SYSTEM: *Israeli officials responsibility of airport safety*

REFERENCES:

Israeli officials are responsible for airport security

Israel is in charge of the safety 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

Clipped N-gram Precision

Clipped N-gram Precision

SYSTEM: the the the the the the the

REFERENCE: The cat is on the mat

What is the unigram precision?

Clipped N-gram Precision

SYSTEM: the the the the the the the

REFERENCE: The cat is on the mat

What is the unigram precision?

Not 7/7 but 2/7

Clipped N-gram Precision

SYSTEM: the the the the the the the

REFERENCE: The cat is on the mat

What is the unigram precision?

Not 7/7 but 2/7

Why?

Clipped N-gram Precision

SYSTEM: the the the the the the the

REFERENCE: The cat is on the mat

What is the unigram precision?

Not 7/7 but 2/7

Why?

Because the number of times “the” counts as a correct match is clipped by the number of times it occurs in the reference

Clipped N-gram Precision

SYSTEM: the the the the the the the

REFERENCE: The cat is on the mat

What is the unigram precision?

Not 7/7 but 2/7

# Correct unigrams	7
# **Clipped** correct unigrams	2

Why?

Because the number of times "the" counts as a correct match is clipped by the number of times it occurs in the reference

BLEU

An artificially high precision can be obtained by minimising the length of the translation

To prevent this, the brevity penalty is used

BLEU

- If length of reference and output equal
  1
- If the output is longer than the reference
  1
- If the output is shorter than the reference
  - less than 1, i.e. lower final score

BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision:
- Brevity Penalty:
- Final Score:

BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision: p1= p2= p3= p4=
- Brevity Penalty:
- Final Score:

BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision: p1=1.0(8/8) p2= p3= p4=
- Brevity Penalty:
- Final Score:

BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision: p1=1.0(8/8) p2=0.86(6/7) p3=0.67(4/6) p4=0.6 (3/5)
- Brevity Penalty:
- Final Score:

BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision: p1=1.0(8/8) p2=0.86(6/7) p3=0.67(4/6) p4=0.6 (3/5)
- Brevity Penalty: c=8, r=9, BP=0.8889
- Final Score:

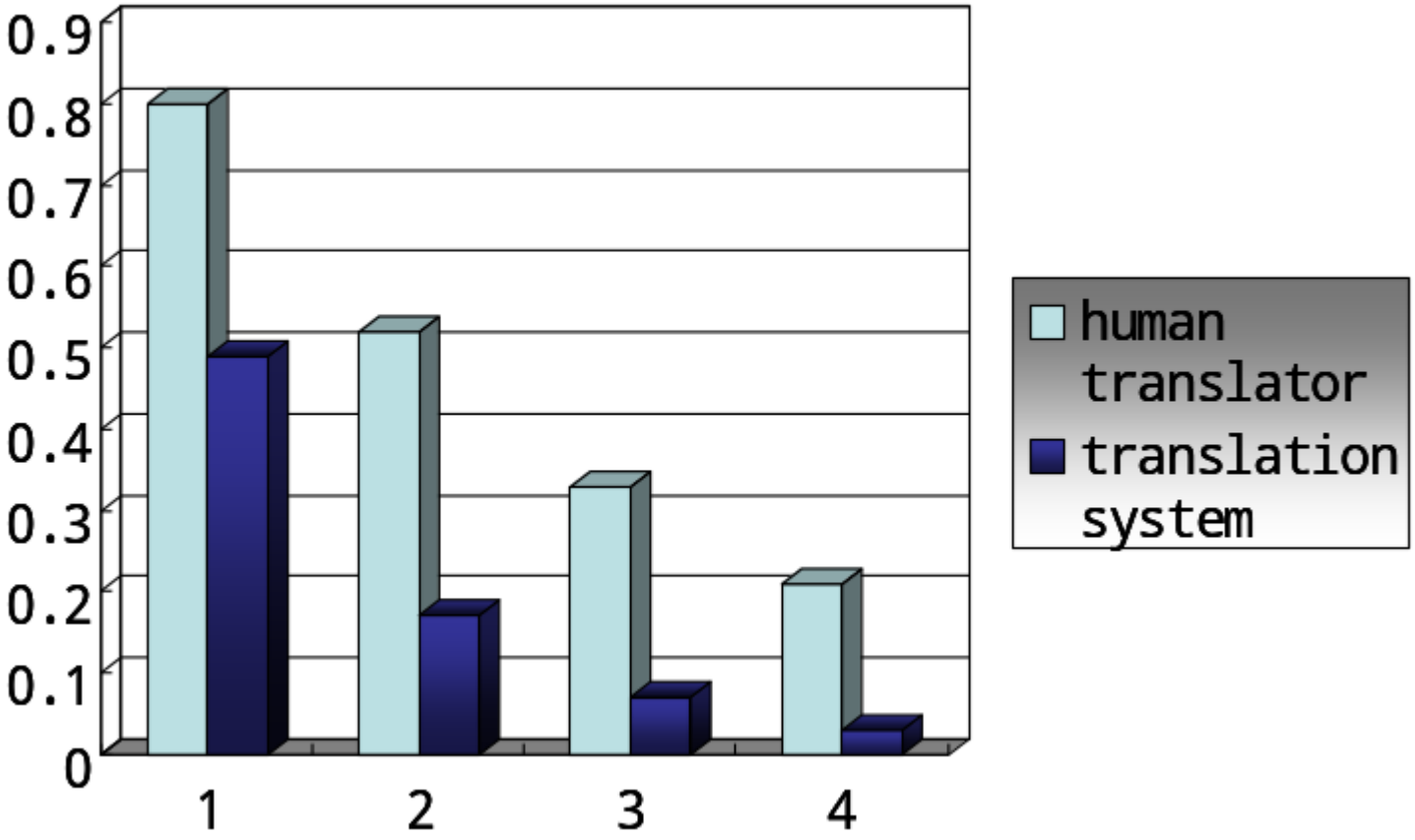
BLEU: an example

MT Hypothesis	The gunman was shot dead by police .
Ref 1	The gunman was shot to death by the police .
Ref 2	The gunman was shot to death by the police .
Ref 3	Police killed the gunman .
Ref 4	The gunman was shot dead by the police .

- Precision: p1=1.0(8/8) p2=0.86(6/7) p3=0.67(4/6) p4=0.6 (3/5)
- Brevity Penalty: c=8, r=9, BP=0.8889
- Final Score:

$$\sqrt[4]{1\times 0.86\times 0.67\times 0.6}\times 0.8889 = 0.6816$$

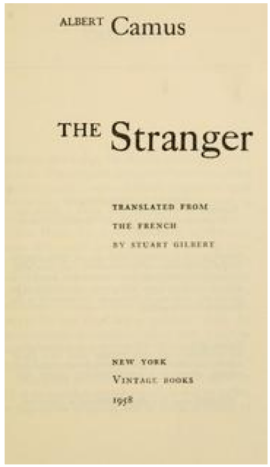
BLEU: Human vs Machine

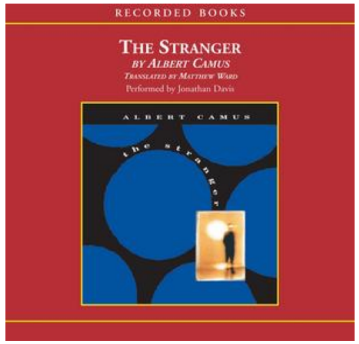


BLEU: Human vs Machine

- Two translations of L'ètranger (Camus) into English, by Gilbert and Ward
- Use Gilbert as reference

System	BLEU
Ward	0.185
Google Translate	0.112





Alternatives to BLEU

- NIST
- METEOR
- And many others...

NIST

- BLEU gives all n-grams equal weight
- NIST calculates how informative a particular n-gram is
- When a correct n-gram is found, the rarer it is, the more weight it will be given
- For example, if the bigram "on the" is correctly matched, it will receive lower weight than a bigram such as "interesting calculations", as the latter is less likely to occur
- NIST values not in the range [0,1]

METEOR

- Partial credit for matching stems

SYSTEM: Jim **went** home

REFERENCE: Joe **goes** home

- Partial credit for matching synonyms

SYSTEM: Jim **walks** home

REFERENCE: Joe **goes** home

- Use of paraphrases

Criticisms of Automatic Metrics

- Ignore relevance of words

(names and core concepts more important than determiners and punctuation)

- Operate on local level

(do not consider overall grammaticality of the sentence or its meaning)

- Scores are meaningless in isolation

(scores very test-set specific, absolute value not informative)

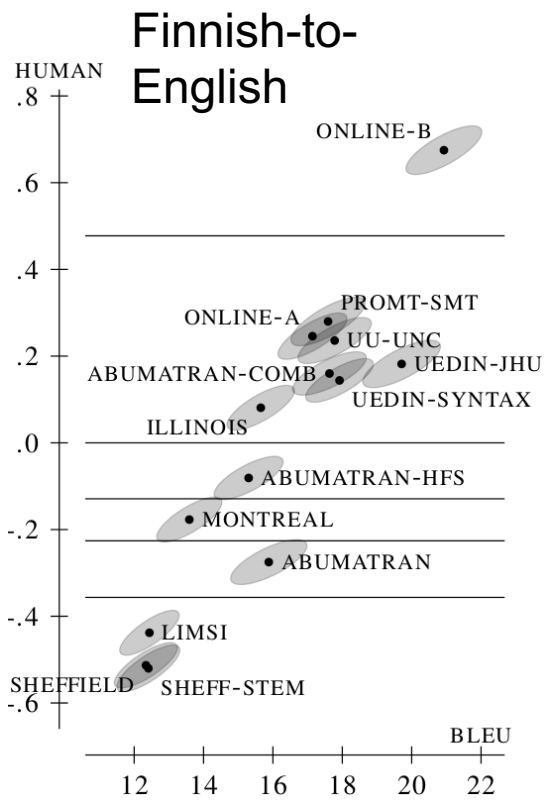
- Human translators score low on BLEU

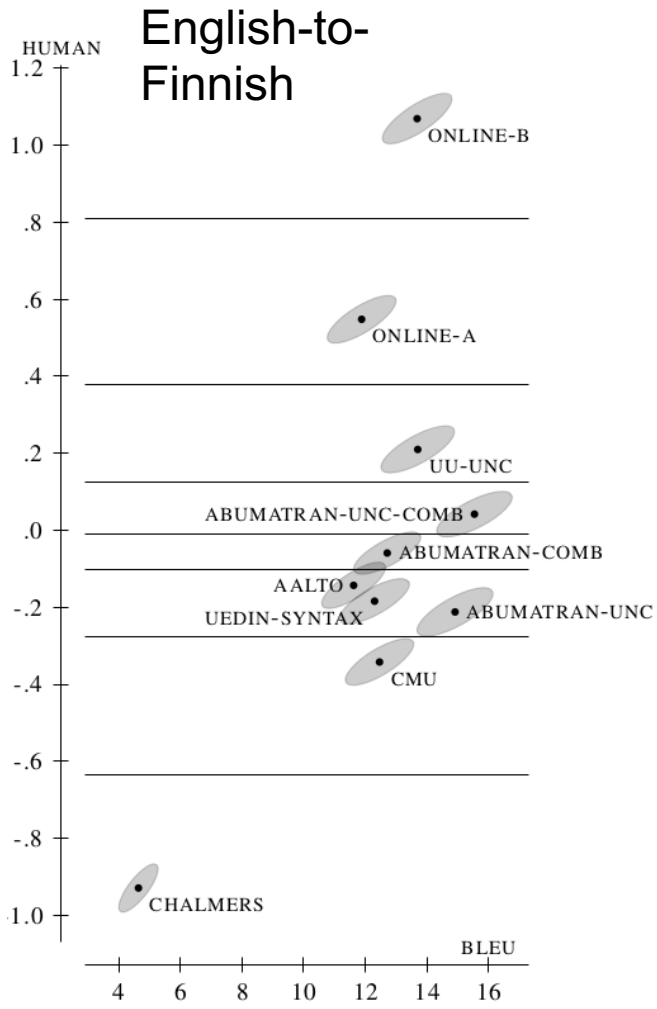
(possibly because of higher variability, different word choices)

Criticisms of BLEU

- Not designed to test individual sentences
- Not meant to compare different MT systems
  - Penalises rule-based systems
- Extremely useful tool for system developers!
- Bad correlation for morphologically-rich languages

BLEU correlation





(WMT 2015)

Why Automatic Evaluation?

- While we develop an MT system, we want to know if performance improves whenever we make any changes
- Cheap, consistent evaluation is necessary for MT research
- The use of automatic evaluation metrics greatly promote the research progress of statistical MT

Content

1. Introduction
2. Human Evaluation
3. Automatic Evaluation
4. **Task-based Evaluation**

Task-based evaluation

- Machine translation is a means to an end
- Does machine translation output help accomplish a task?

Example Tasks

1. producing high-quality translations by post-editing MT output (MT for publishing)
2. information gathering from foreign language sources (MT for gisting)

Post-editing effort

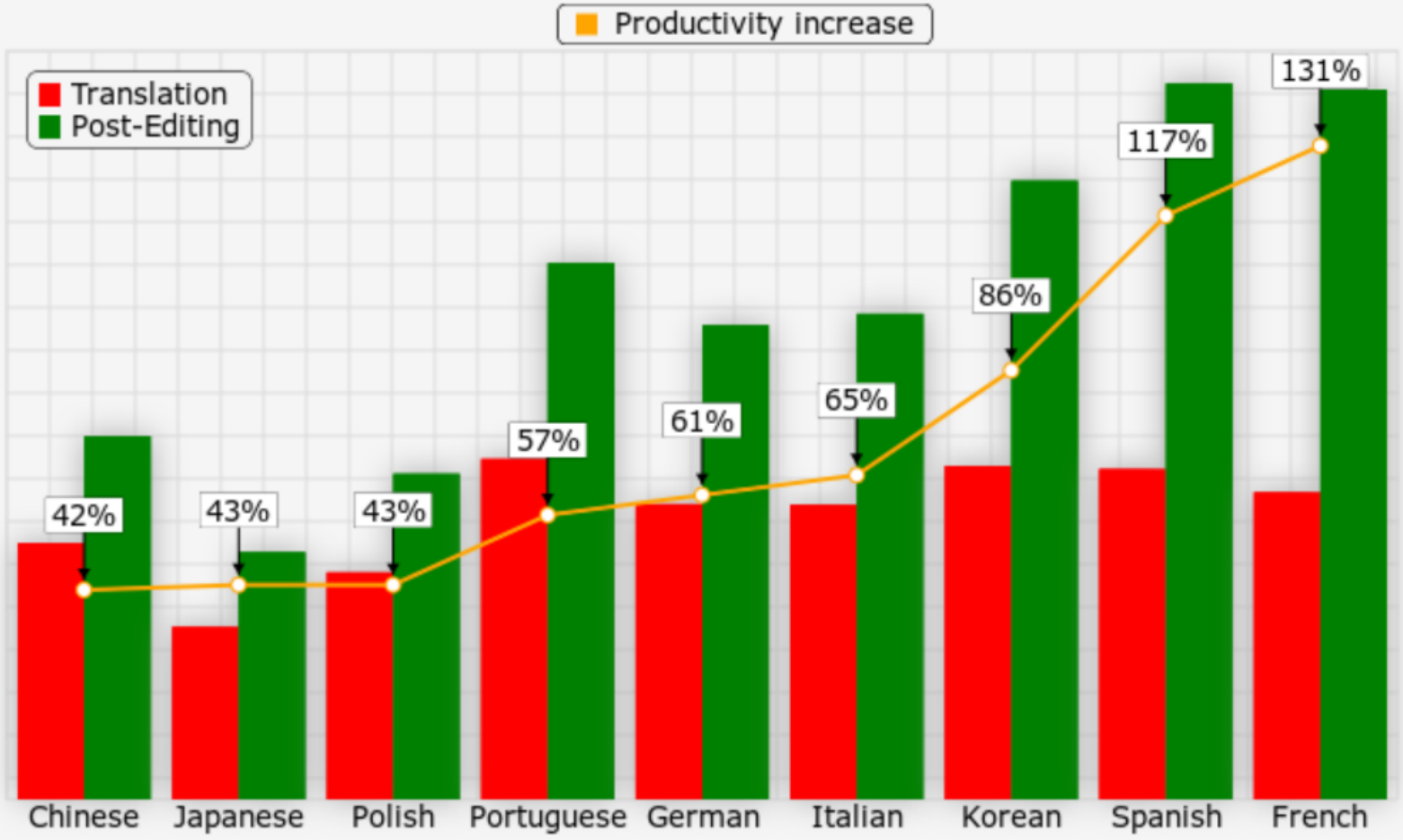
Measuring time spent on producing translations

- baseline: translation from scratch
- post-editing machine translation

But: time depends not only on MT quality, also on post-editor's skills!

Productivity per Language – Translation vs Post-Editing

For all languages tested – in fact for all 37 test participants –, post-editing productivity was significantly higher than translation productivity.



Content Understanding

Given MT output, can monolingual speakers (target language) answer questions about it?

1. Basic facts: who? where? when? names, numbers, and dates
2. Actors and events: relationships, temporal and causal order
3. Nuance and author intent: emphasis and subtext

Content Understanding

- Sentence editing task (WMT 2009-2010)
- person A edits the translation to make it fluent (with no access to source or reference)
- person B checks if edit is correct

Did person A understand the translation correctly?

Discussion

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computer Science

http://www.cit.ie

Thank you

Haithem. afli@cit.ie

@AfliHaithem