CA4012

Statistical Machine Translation

Week 3: MT Evaluation

Lecturer: Dr. Sheila Castilho

2nd Semester 2019-2020

Recap and Quiz

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

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

Recap and Quiz

- What is the difference between fluency
- and adequacy?

Content

1. **Introduction**
2. Human Evaluation
3. Automatic Evaluation
4. Task-based Evaluation
5. Diagnosis and Estimation

Why do we need evaluation in MT?

- Evaluation provide data on whether a system works and why, which parts of it are effective and which need improvement.

- Evaluation needs to be **honest** and **replicable**, and its methods should be as rigorous as possible.

First question: what is quality?

- Evaluation is a complex problem

- What does quality mean?
  - Fluent? Adequate? Both? Easy to post-edit? Usable? All of them? None of them?

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.

(From 2001 NIST evaluation)

Quality for whom/what?

- Why are you evaluating the MT system?
  - End-user (gisting vs dissemination)
  - Post-editor (light vs full post-editing)
  - Other applications (e.g. Cross Lingual IR)
  - MT-system (tuning or diagnosis for improvement)

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

How to evaluate MT?

- A few methods
- Automatic evaluation
  - Automatic evaluation metric (AEMs)
  - Automatic classifications

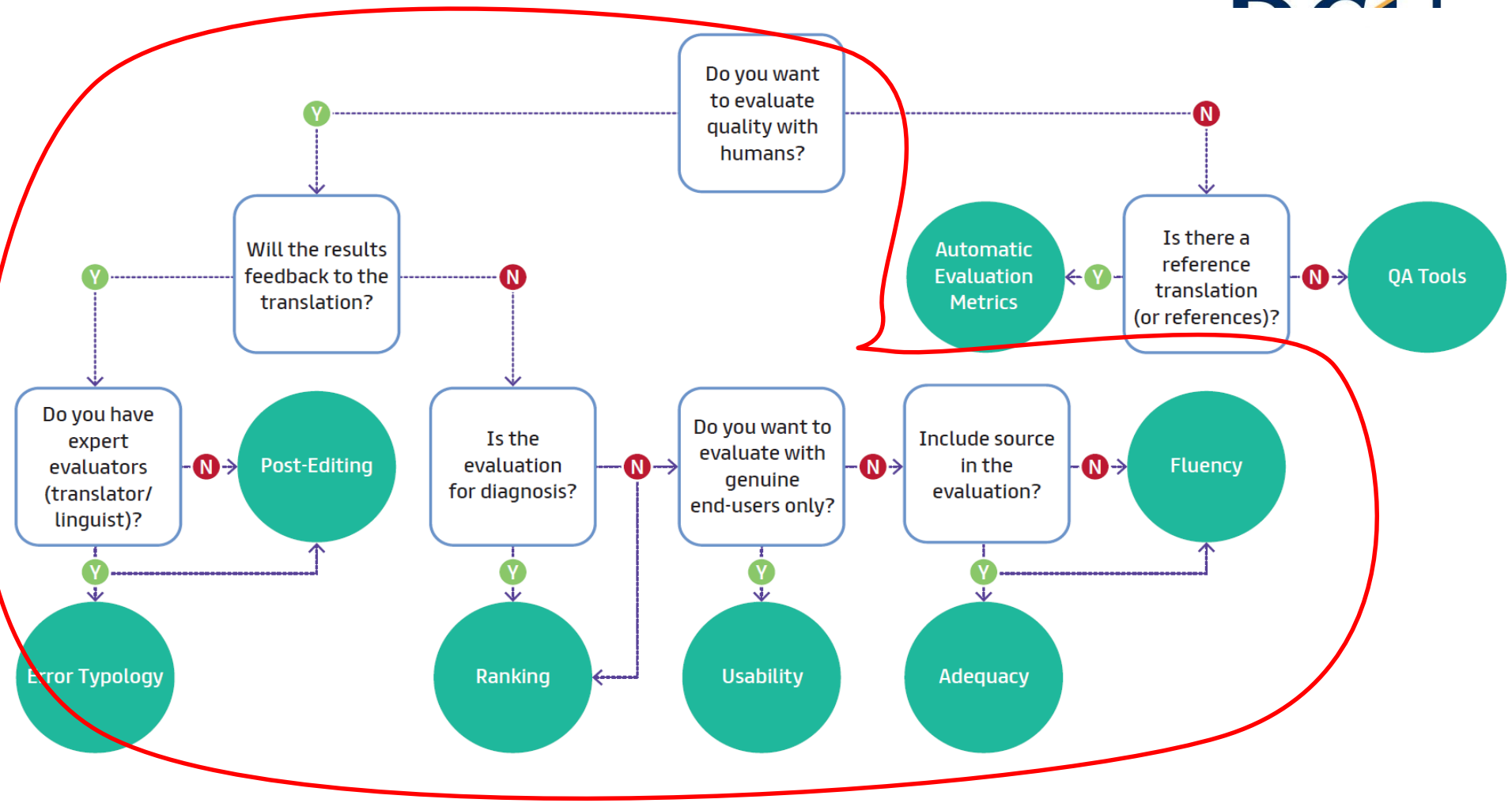
- Human Evaluation
  - Human evaluation metrics (HEMs)
  - Professional translators/bilinguals/crowd
  - User Evaluation
    - Usability, reading comprehension (UEMs)

How to evaluate MT?

- Quality Assessment tools (QA)
  - (Semi)automatic
  - Heavily (still) applied in industry

- Quality Estimation (QE)
  - “Not really evaluation”
  - reference translation is not available

Translation Evaluation Flowchart



Moorkens, J., Castilho, S., Gaspari, F., Doherty, S (Ed.). (2018) *Translation Quality Assessment: From Principles to Practice.* Heidelberg: Springer.

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**

Adequacy

- also known as “accuracy” or “fidelity”
- Focus on the **source** text

“the extent to which the translation transfers the meaning of the source text translation unit into the target”

- Likert scale:
    1. None of it
    2. Little of it
    3. Most of it
    4. All of it

- Why is Adequacy useful for MT evaluation?
  - It tells us how much of the source message has been transferred to the translation
  - Sometimes you are only interest in the meaning of the source sentence

Fluency

- also known as intelligibility
- focuses on the target text
- “the flow and naturalness of the target text unit in the context of the target audience and its linguistic and sociocultural norms in the given context”
- Likert scale:

1. No fluency
2. Little fluency
3. Near native
4. Native

- Why is Fluency useful for MT evaluation?
  - It tells if the message is fluent/intelligible (i.e. sounds natural to a native speaker) or if it is “broken language”.

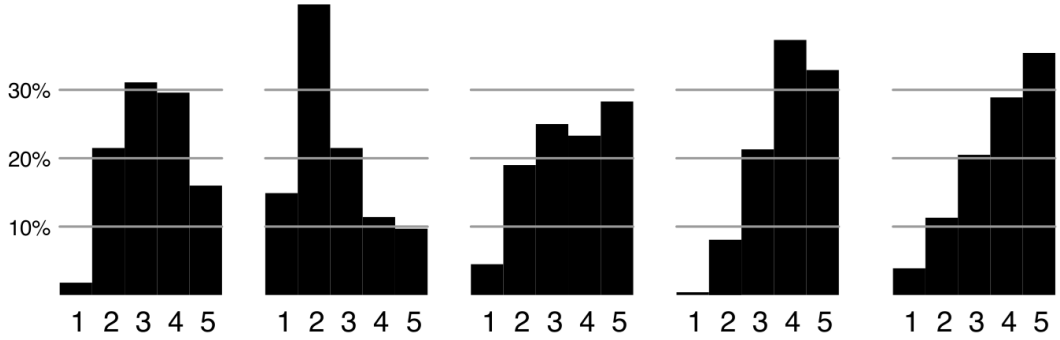
Adequacy-Fluency

- Adequacy and Fluency generally go together
  - But sometimes you may want to prioritise one over the other
    - Technical documentation may require more adequacy

Human Evaluation

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

- Still, evaluators disagree (distributions)



(WMT 2006 evaluation task)

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

- 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)?

Inter-annotator agreement (IAA)

Necessary because human annotation/evaluation is:

- Subjective
- Prone to errors (fatigue)
- Biased (preference for a label)
- Based on human-written guidelines (misinterpreted)

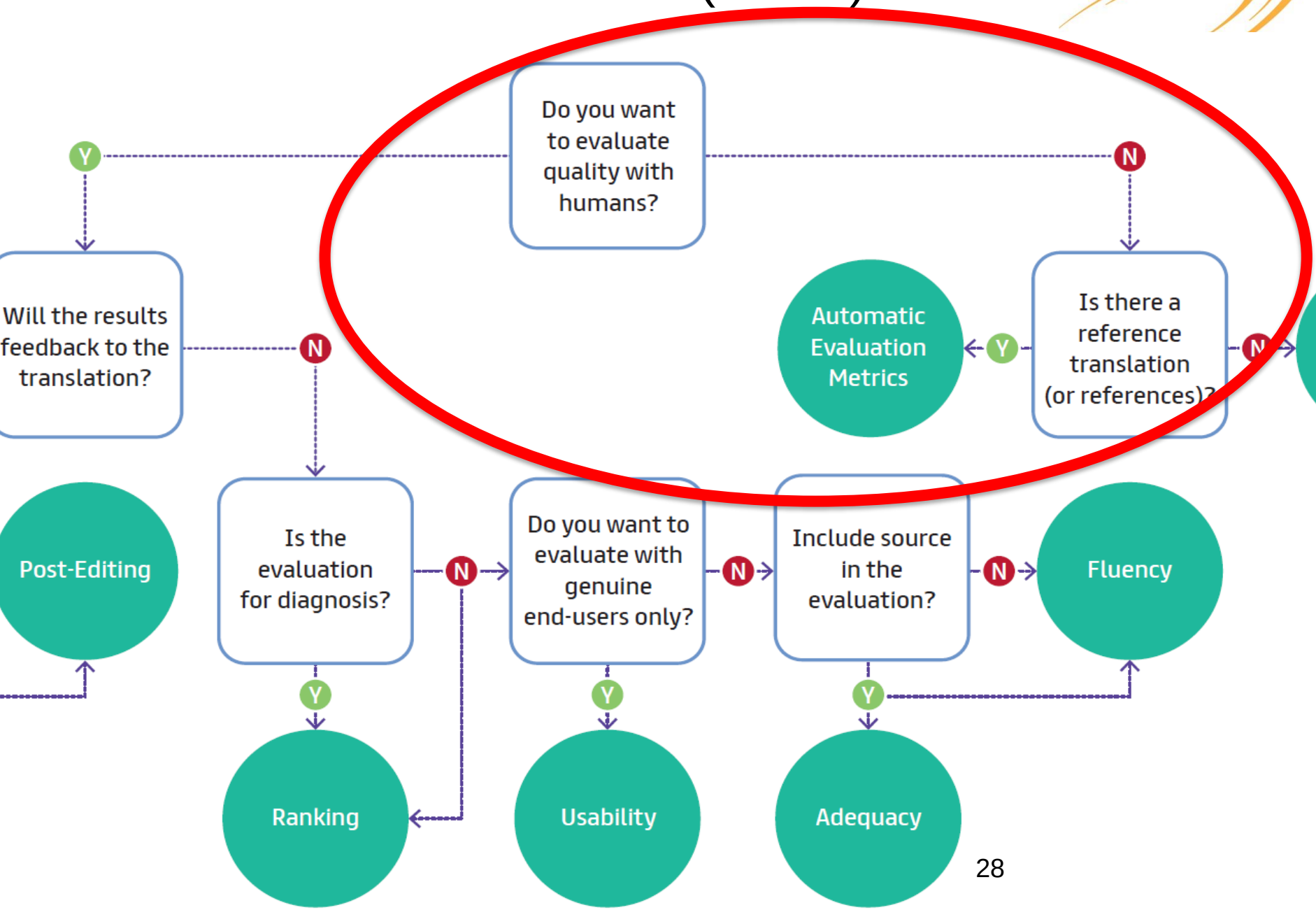
Results can:

- Identify improvements needed in annotation scheme
- Indicate usefulness of data
- Indicate replicability of data (e.g. clinical diagnoses)

Most used coefficient:

- Cohen's Kappa (weighted and non-weighted)
- Fleiss' Kappa

Automatic Evaluation Metrics (AEMs)



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?

- How many insertions?
- How many substitutions?
- How many deletions?

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

WER Calculation

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

WER Calculation

reference translation

MT output

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

WER Calculation

reference translation

MT output

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

Label all matched words

WER Calculation

reference translation

MT output

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

WER Calculation

reference translation

MT output

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

Search a shortest path from top-left corner to bottom-right corner (dynamic programming)

WER Calculation

reference translation

MT output

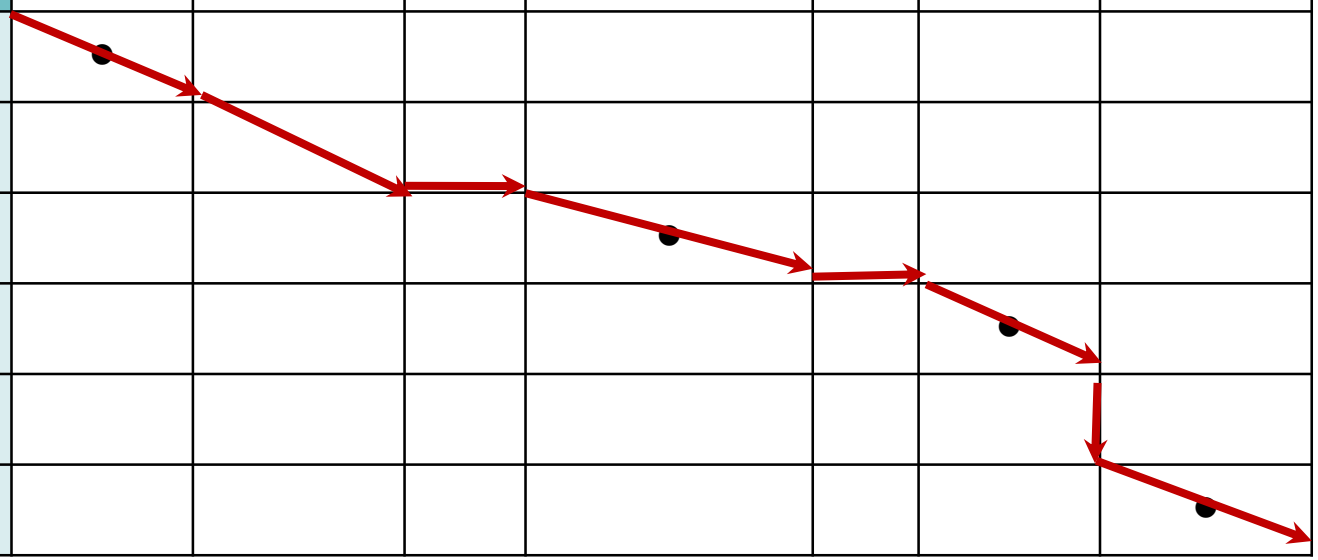
	Israeli	officials	are	responsible	for	airport	security
Israeli							
official							
responsible							
airport							
is							
security							

WER Calculation

reference translation

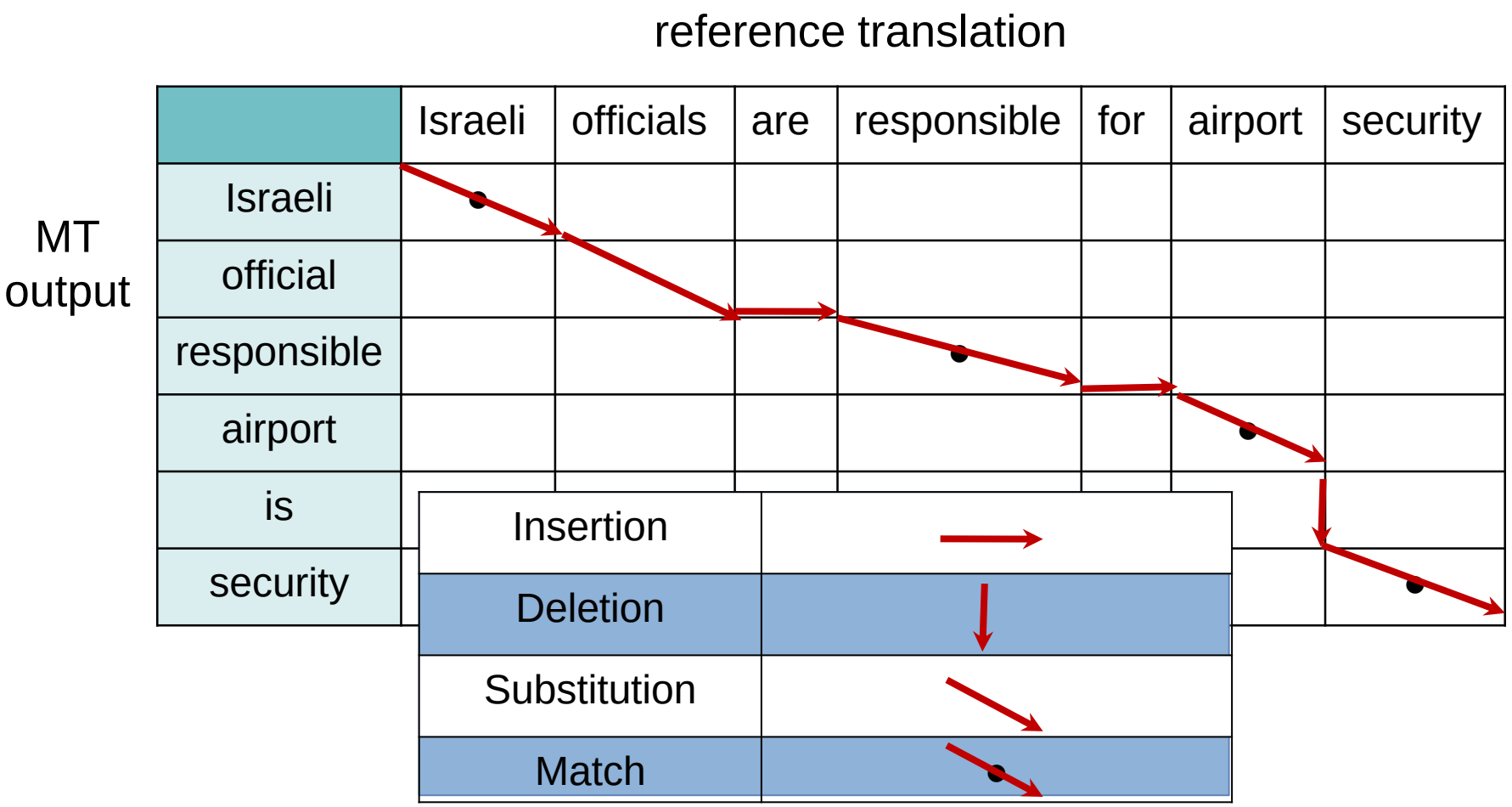
MT output

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



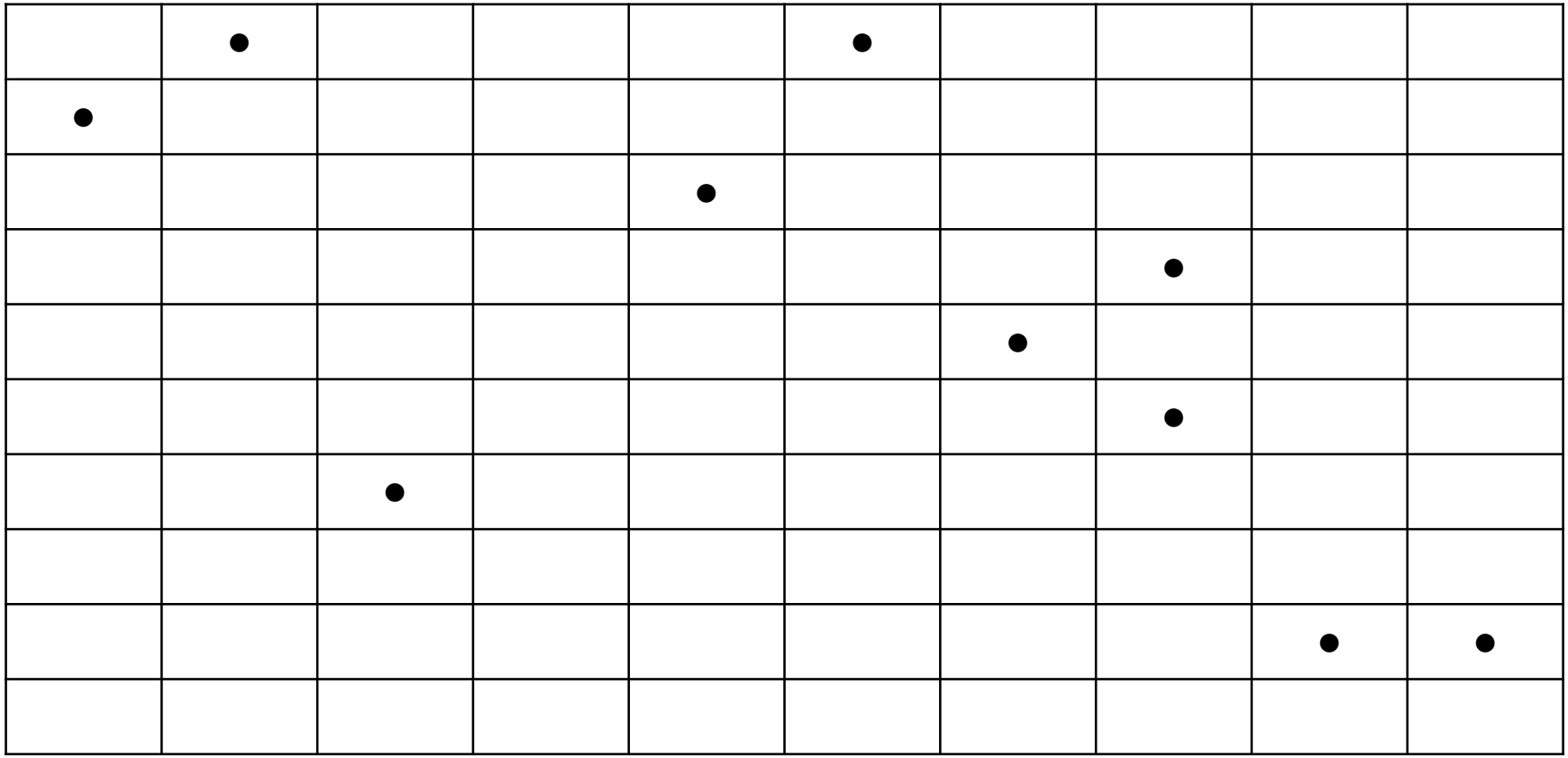
Calculate the number of different operations

WER Calculation



Problem

Consider a more complex case



Task. Design an algorithm to search the shortest path

Problem

Consider a more complex case

reference translation

MT output		Israeli	officials	are	responsible	for	airport	security
	0	1	2	3	4	5	6	7
Airport	1	1	2	3	4	5	5	6
security	2	2	2	3	4	5	6	5
Israeli	3	2	3	3	4	5	6	6
officials	4	3	2	3	4	5	6	7
are	5	4	3	2	3	4	5	6
responsible	6	5	4	3	2	3	4	5

Israeli officials are responsible for airport security

Airport security Israeli officials are responsible for airport security

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 -> **high WER score**

Translation Error Rate (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.

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

Translation Error Rate (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)

Translation Error Rate (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
```

Example from (Snover et al. 2006)

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”

6 1-grams (or unigrams)

-

5 2-grams (or bigrams)

-

4 3-grams (or trigrams)

-

3 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"

6 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

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

- BLEU doesn't care about deletions, substitutions, insertions, just matches, and on the phrase-level.
- N-gram precision is the ratio of correct n-grams of a certain order n in relation to the total number of generated n-grams of that order.
- The brevity penalty reduces the score if the MT output is too short.

Clipped N-gram Precision

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

$$Precision_n = \frac{number\ of\ clipped\ correct\ ngram\ in\ output}{total\ number\ of\ ngram\ in\ output}$$

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

- If any n-gram precision is 0 => BLEU=0
- BLEU is commonly computed over entire test set.

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

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

$$\min(1, \frac{output\ length}{reference\ length})$$

BLEU

$$\min(1, \frac{output\ length}{reference\ length})$$

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

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
  - Penalises NMT systems
  - Why???
- Extremely useful tool for system developers!
- Bad correlation for morphologically-rich languages

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 MT (satistical, neural).

Content

1. Introduction
2. Human Evaluation
3. Automatic Evaluation
4. **Task-based Evaluation**
5. Diagnosis and Estimation

Task-based evaluation

- Machine translation is a means to an end product – high quality translation.
- 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)

Pos-editing (PE)

- The “term used for the correction of machine translation output by human linguists/editors” (Veale and Way 1997)
- “checking, proof-reading and revising translations carried out by any kind of translating automaton”. (Gouadec 2007)
- Common use of MT in production – over 80% of Language Service Providers now offer post-edited MT (Common Sense Advisory 2016)

Measurement of PE effort

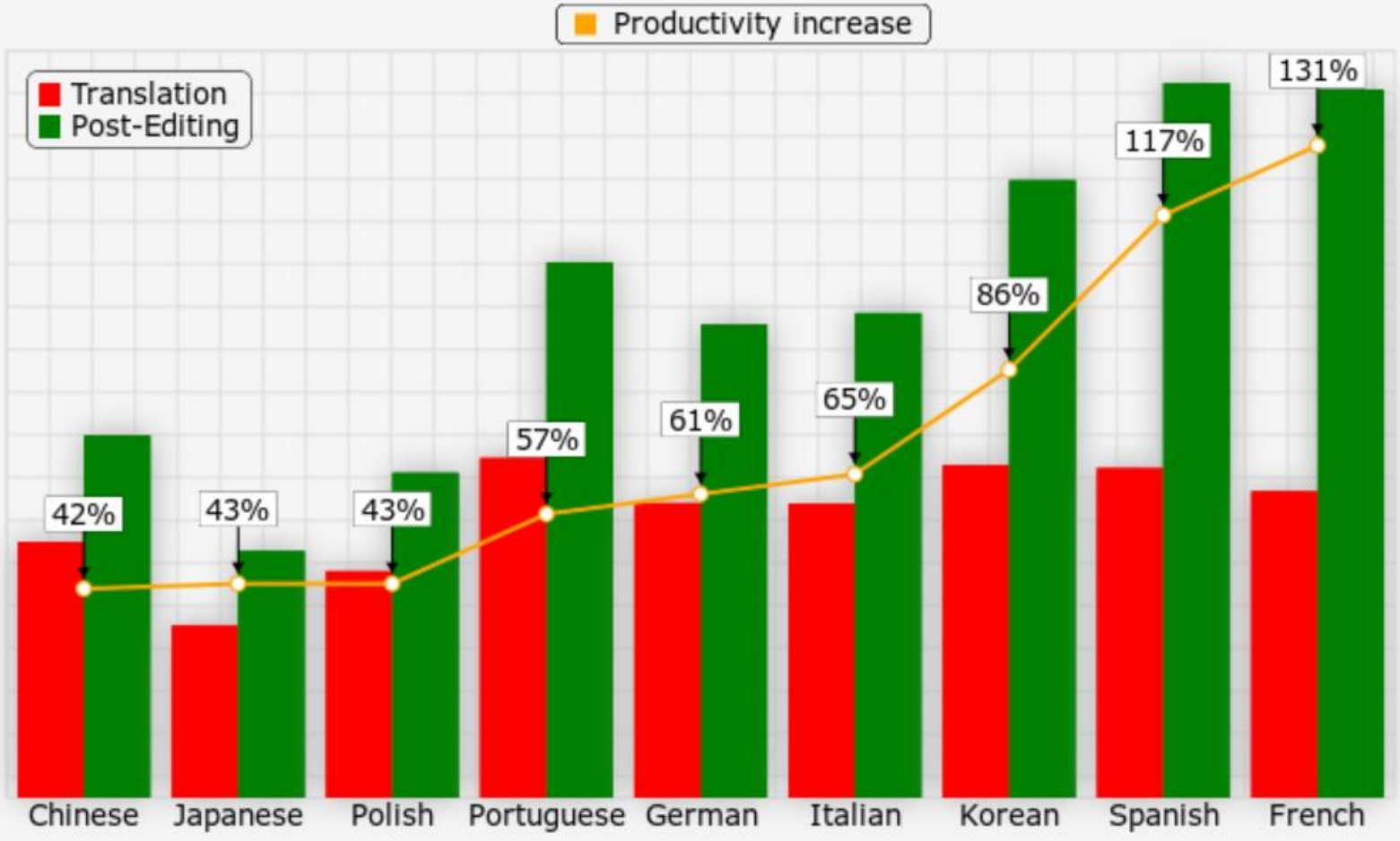
From Krings' book Repairing Texts (2001)

- Temporal effort
  - Throughput, the amount of time spent post-editing
  - Often expressed in words/second

- For MT Eval – faster better means better MT output?
  - productivity

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.



Measurement of PE effort

- Technical effort
  - The number of edit operations made
  - Often approximated using hTER automatic metric

  - For MTEval – fewer edits mean better MT
    - Correlates with time effort = productivity

  - HTER
    - PE as reference
    - PE as hypothesis

Measurement of PE effort

- Cognitive effort
  - May be measured in several ways
  - In DCU we often use eye-tracking

  - For MT Eval – less cognitive effort means better MT output
  - Cognitive effort has been correlated to other HEMs

PE

- Why use post-editing for Machine Translation evaluation?

  - Assess usefulness of MT system in production
  - Identify common errors
  - Create new training or test data

  - However, measurements of post-editing effort tend to differ between novice (students) and professionals, and crowd and professionals

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?

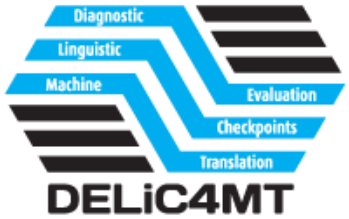
Diagnostics Evaluation

- In a diagnostics evaluation, the translation quality is not assessed overall, instead, it is assessed on a set of predefined **check points**
- The check points are defined according to a linguistic category taxonomy

Diagnostics Evaluation

Chinese			English		
Word level			Word level		
Ambiguous word	New word	Idiom	Noun	Verb (with tense)	Modal verb
Overlapping word	Collocation	Noun	Adjective	Adverb	Pronoun
Verb	Adjective	Adverb	Preposition	Ambiguous word	Plurality
Pronoun	Preposition	Quantifier	Possessive	Comparative and superlative degree	
Phrase level			Phrase level		
Subject-predicate phrase	Predicate-object phrase	Preposition-object phrase	Noun phrase	Verb phrase	Adjective phrase
Measure phrase	Location phrase		Adverb phrase	Preposition phrase	
Sentence level			Sentence level		
BA sentence[2]	BEI sentence[3]	SHI sentence	Time clause	Reason clause	Condition clause
YOU sentence	Compound sentence		Result clause	Purpose clause	

Diagnostics Evaluation



OUTPUT

Sentence	6
Checkpoint source	mr.
Checkpoint target	monsieur
Alignment	4-4
Source sentence	let us remember , Mr. Speaker , that these segments of our society form the backbone of our economy .
Reference	Target sentence: souvenons - nous , monsieur le Orateur, que ce sont ces secteurs de notre Societé qui servent de épine dorsale à notre économie .
MT output	Souvenons-nous , Mr. l' orateur , que ces segments de notre société constituent l' épine dorsale de notre économie .
Checkpoint ngrams	0/1

Sentence	6
Checkpoint source	speaker
Checkpoint target	orateur
Alignment	5-6
Source sentence	let us remember , Mr. Speaker , that these segments of our society form the backbone of our economy .
Reference	Target sentence: souvenons - nous , monsieur le Orateur , que ce sont ces secteurs de notre Societé qui servent de épine dorsale à notre économie .
MT output	Souvenons-nous , Mr. l' orateur , que ces segments de notre société constituent l' épine dorsale de notre économie .
Checkpoint ngrams	1/1

Overall ngrams	803/1598
Overall recall	0.50250316
Brevity penalty	1.0
Overall score	0.50250316

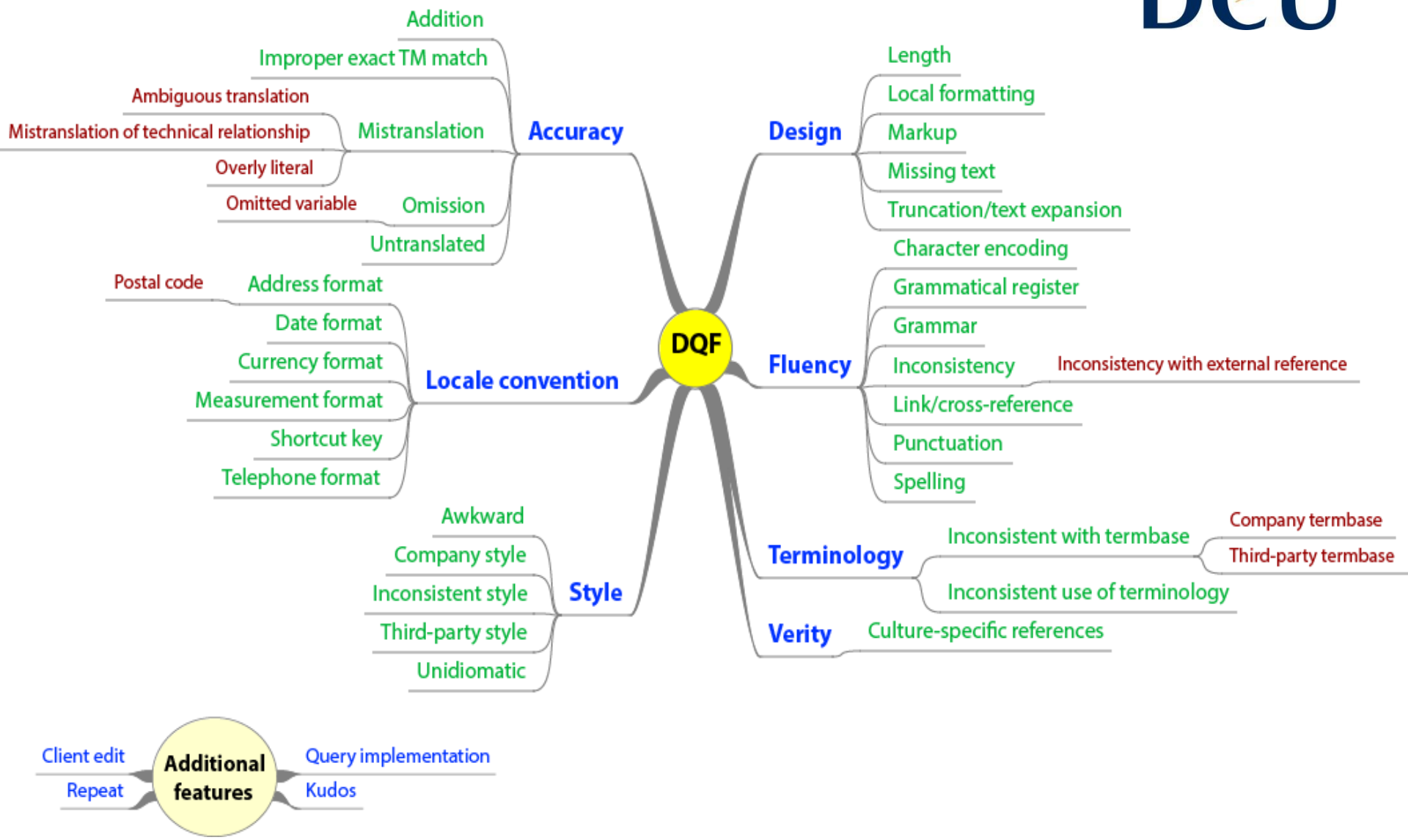
Error Classification

- Identify and classify errors in a translated text
- A few taxonomies have been proposed

- Vilar et al. (2006)
- Llitjós et al. (2005).
- Federico et al. (2014)
- Costa et al. (2015)
  - DQF – TAUS
  - MQM – QT21[2]
    - Harmonized

[2] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

DQF / Multidimensional Quality Metrics



Error Classification

- Why use error taxonomies for translation evaluation?
  - Identify types of errors in MT or human translation
  - Detailed error report is useful for adjusting MT systems, reporting back to clients
  - LSPs use taxonomies and severity ratings to monitor translators' work

Error Classification

- More possible analyses:
  - relations between particular error types and user/post-editor preferences
  - the impact of different error types on different aspects of post-editing effort

    - *However, error annotation is expensive*
  - Automatic Error Classification has been Proposed (See Popovic 2018)

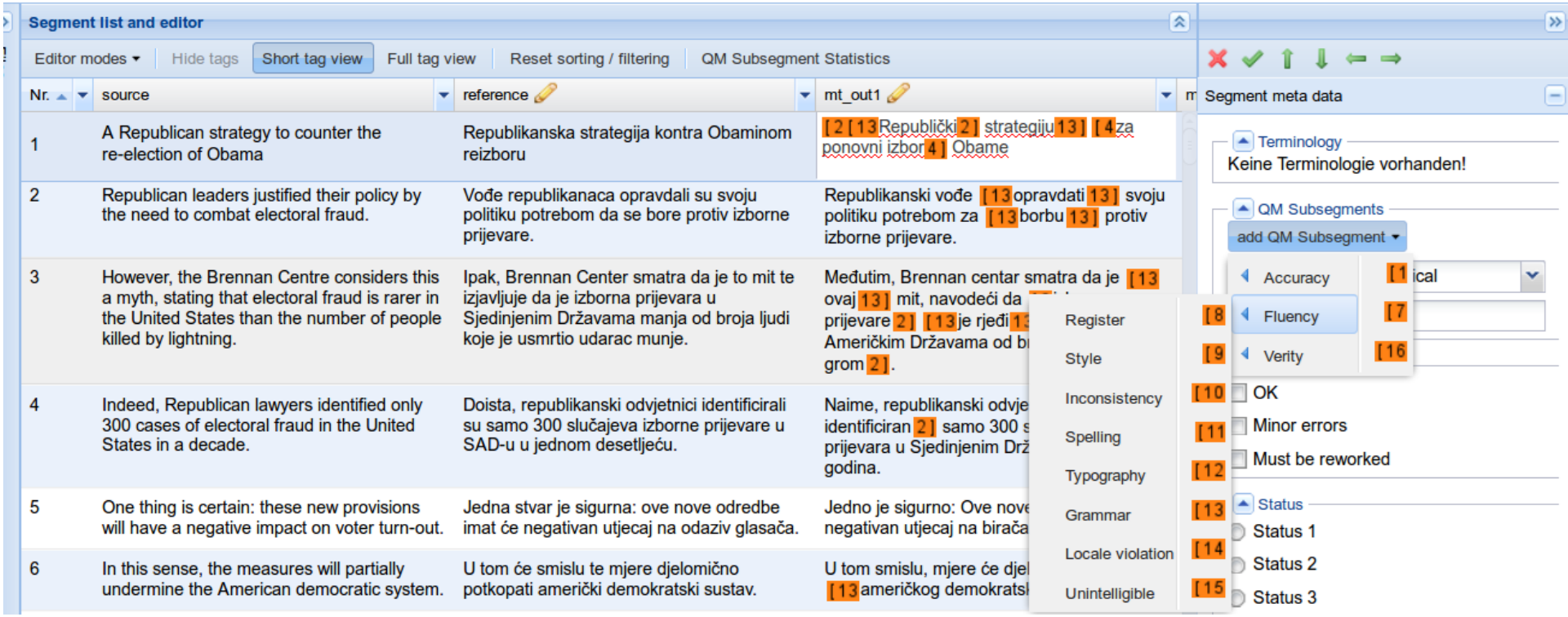
MQM

translate5

DCU

Authenticated user: Project Manager
Login name: manager
Task: m3_enhr

Logout | Tasks | My settings | English



What about Quality Estimation (QE)?

- Not a measure of quality: no comparison against a reference
- Provides an **estimate** on the quality of translations on the fly

Quality is data driven:

- Can the translation be published as it is?
- Can a reader get the gist?
- Is it worth post-editing it?
- How much effort to fix it?

What about Quality Estimation (QE)?

- Features extracted from examples of translation and source
  - Source -> complexity features (i.e. how hard it is to translate?): sentence length, common words (frequency of words)
  - Translation-> fluency features: grammatical (i.e. grammar checker), sequence of words
  - Source+ translation -> adequacy features (i.e. difference in length)
  - PEs and human annotated data can also be used

What about Quality Estimation (QE)?

- Features can be tailored and extracted depending on the definition of quality
  - How much effort to fix the translation?
    - PE time
  - Can the translation be published as it is?
    - Adequacy, fluency scores
  - Can a reader get the gist?
    - General adequacy, fluency from sample translations

References

- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation". ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.
- Callison-Burch, C., Osborne, M. and Koehn, P. (2006) “Re-evaluating the role of BLEU in Machine Translation Research" in 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006 pp. 249–256
- Antonio Toral, Sudip Kumar Naskar, Federico Gaspari, Declan Groves. DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. The Prague Bulletin of Mathematical Linguistics No 98, 2012, pp. 121-131., ISSN 0032-6585, DOI: 10.2478/v10108-012-0014-9
- Zhou, Ming, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. Diagnostic evaluation of machine translation systems using automatically constructed linguistic checkpoints. In *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1*, COLING ’08, pages 1121–1128, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.

Exercise

Calculate BLEU unigram, bigram, trigram and 4-gram precision as well as the brevity penalty for the following three translations:

- **Translation 1:** *Salmons swim in river .*
- **Translation 2:** *Fish swim in the river .*
- **Translation 3:** *The salmon swam in the river .*
- **Reference:** *Salmons swim in the river .*

Discussion

Acknowledgement

Parts of the content of this lecture are taken from previous lectures and presentations given by Qun Liu, Jennifer Foster, Declan Groves, Yvette Graham, Kevin Knight, Josef van Genabith and Andy Way.