CA4012

Machine Translation

Week 3: MT Evaluation

Lecturer: Dr. Sheila Castilho

2nd Semester 2021-2022

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?

Why do we need evaluation in MT?

- Evaluation provides information on whether an MT 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.

MT evaluation: not a trivial task

- there are many ways to say the same thing
- there are many ways to translate something into another language

⇒ There is no single correct translation of a given text

Furthermore:

- are all possible translations equally good?
- what does “good” actually mean in context of translation?

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)

What is a “good translation”?

- Fluent?
- Adequate?
- Both?
- Understandable?
- Easy to correct (post-edit)?
- Usable?
- All of it?
- None of it?
- Something completely different?

Good for whom/what?

- Who/what is the MT system for?
  - end user (gisting vs dissemination)
  - post-editor (light vs full post-editing)
  - other applications (e.g. Cross Lingual IR, multilingual text classification, etc.)
  - MT-system developer

    (tuning or diagnosis for improvement)

Goals for MT Evaluation

- Meaningful: result should give intuitive interpretation of translation quality
- Consistent: repeated evaluation 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 Aspects

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

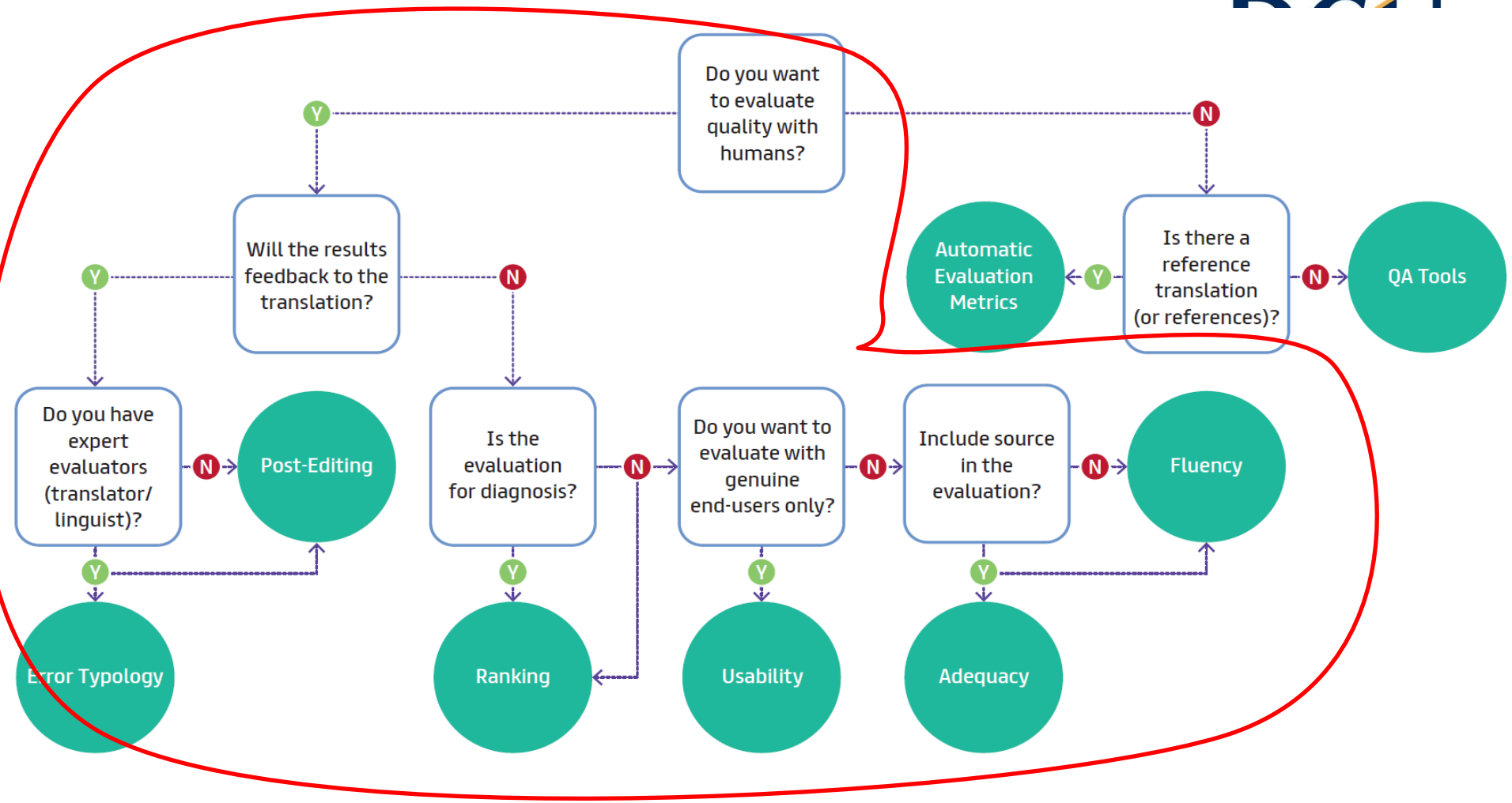
- Automatic evaluation
  - Automatic evaluation metric (AEMs)
  - Automatic classifications
- Human (manual) Evaluation
  - Human evaluation metrics (HEMs)
  - User Evaluation
    - Usability, reading comprehension (UEMs)
  - Professional translators/bilinguals/crowd

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 text and/or reference translation
    - **Reference translation**: a translation produced by a trained translator (human)
- Task: **human evaluator asesses the quality of MT output**

Human Evaluation

English-to-Irish example

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

What is “quality”?

Based on at least one of the following criteria:

- adequacy
  - is the meaning of the translation the same as the meaning of the source text?
- fluency
  - is the translation grammatically correct or broken language?
- comprehensibility
  - is it clear what the translation means?

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”

- Might be assessed on a Likert scale, i.e:

1. None of it
2. Little of it
3. Most of it
4. All of it

or

1. None of it
2. Little of it
3. Some of it
4. Most of it
5. 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”
- Might be assessed on a Likert scale, i.e:

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

or

1. No fluency
2. Little fluency
3. Some fluency
4. Near Native
5. 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

Comprehensibility

- also known as “comprehension”, “readability”
- focusses **exclusively** on the **target** text

tells if the translation can be understood

- Used much less than adequacy and fluency
  probably because it is difficult to assess it together with adequacy

note: comprehensible but inadequate translations are **not** good translations

Methods for human evaluation

- assigning a score to each sentence according to the given criterion (adequacy, fluency, …)
  higher score -> better quality

- ranking two or more translations
  from best to worst
  higher rank -> better quality

- error classification
  mark each error in the translation and assign it a type/class

Scoring

Assigning a score is the mostly used method for manual MT evaluation

Or, a continuous value (usually between 0 and 100)

Annotation tool for scoring

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

Ranking

- comparing two or more translations

  is translation A better than translation B,
  or B is better than A,
  or they are equal?

- + simpler for annotators than scoring
  (especially if comparing only two translations)
- - information about overall quality is lost

  A better than B ⇏ A is good and B is bad!

Annotation tool for ranking

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)

Error classification

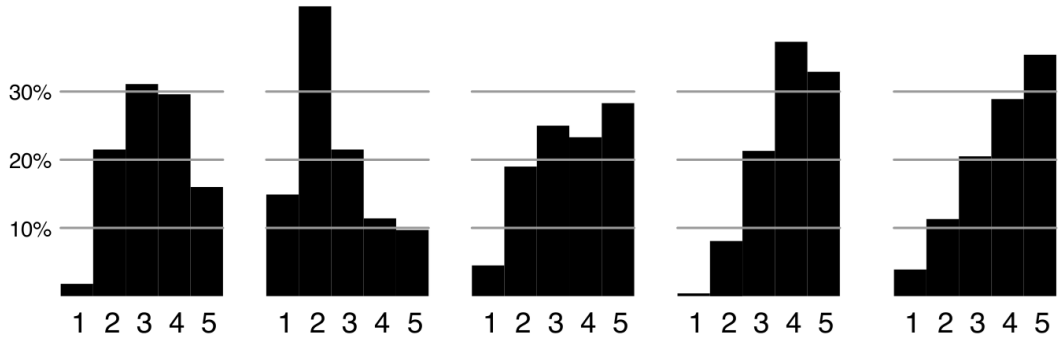
- annotating each error in the translated text
- assigning it a type/class from a pre-defined error scheme (such as “word form”, “lexical choice”, “word order”, etc.)

- + detailed information about the problems is provided
- - requires more time and effort from the evaluators

Human evaluation is subjective

evaluators generally disagree

distributions of scores:



(WMT 2006 evaluation task)

Measuring subjectivity: inter-annotator agreement (IAA)

Human annotation/evaluation is:

- Prone to errors (fatigue)
- Biased (personal preferences)
- Based on human-written guidelines (which can be not enough precise/clear, or misinterpreted)

IAA can:

- Identify improvements needed in evaluation process (including guidelines)
- Indicate usefulness of data
- Indicate replicability of data

Most used metrics:

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

(Dis)advantages of human evaluation

- Meaningful: result should give intuitive interpretation of translation quality **YES!** 
- Consistent: repeated evaluation should lead to same results **SUBJECTIVITY** 
- Correct: metric must rank better systems higher **YES!** 
- Low cost: reduce time and money spent to carry out evaluation **HMMMM….** 
- Tunable: automatically optimise system performance towards metric **see above (and imagine evaluating tenths or hundreds of different versions of a system manually…)** 

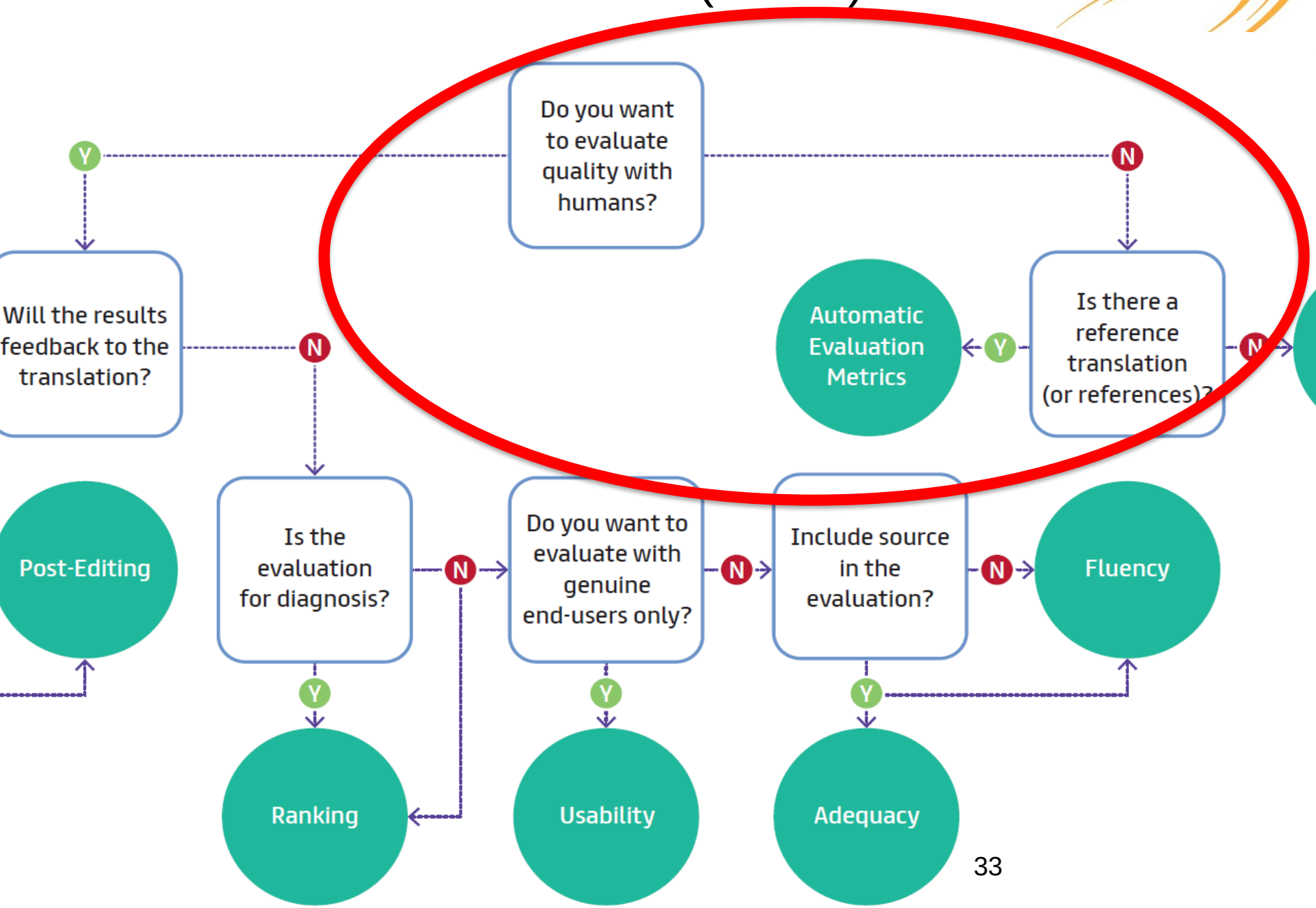
Solution: Automatic Evaluation

automatic evaluation metric:
a computer program which calculates a score related to the translation quality

Advantages

- consistent 
- low cost (both time and money)  
- tunable 

Automatic Evaluation Metrics (AEMs)



Automatic Evaluation

Basic strategy

- Input 1: MT output
- Input 2: human reference translation
- Output: a score which represents the similarity between the MT output and the human reference

What about…

- meaningful 🤔

  What does an automatic score represent? Adequacy? Fluency? Comprehension? Something else?

- correct 🤔

  Will an automatic metric really rank different MT outputs from worst to best?

What does the score represent?

- The score does not represent any translation quality criterion
  - only the similarity between MT output and one of many possible human translations
- This is true for all currently widely used automatic metrics.

- Some recently proposed new metrics try to model translation quality criteria (adequacy, fluency, error count)
  - however, they need data and training

Correctness of automatic evaluation

How to assess whether an automatic evaluation metric is correct?

- if it ranks different MT outputs in the same order as human evaluation scores

  (correlation)

- there are many automatic metrics, and new ones are constantly coming out
- not all of them are equally correct

(Dis)advantages of automatic evaluation

- despite their disadvantages, automatic evaluation metrics greatly promote the research progress of MT

- automatic evaluation metrics are evaluated by comparison with human evaluation scores

  human evaluation remains the “gold standard”

Two main principles of automatic evaluation

- similarity (matching)
- difference (edit distance)

between MT output and human reference translation

Matching: precision and recall

- precision:

  shows how many instances in the *generated set* can be found in the *reference set *(and therefore considered as correct)

- recall:

  shows how many instances in the *reference set* are covered by the *generated set*

for MT:

- reference set = reference human translation
- generated set = MT output
- instances = word sequences (*n*-grams)

precision and recall for MT

$$precision = \frac{matches\ in\ translation\ output}{instances\ in\ translation\ output}$$

$$recall = \frac{matches\ in\ reference\ translation}{instances\ in\ reference\ translation}$$

n-grams

an *n*-gram is a sequence of *n* words

n = order of *n*-gram

example for *n*-grams

"The cat sat on the mat"

six 1-grams (or unigrams)

five 2-grams (or bigrams)

four 3-grams (or trigrams)

three 4-grams

two 5-grams

one 6-gram

"The cat sat on the mat"

six 1-grams (or unigrams)

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

five 2-grams (or bigrams)

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

four 3-grams (or trigrams)

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

three 4-grams

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

two 5-grams

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

one 6-gram

- The cat sat on the mat

1-gram precision and recall

Reference: **It will be considered as** a **sort of bridge .**

MT output: **It will sort of bridge be considered as .**

matches (bold): 9 (including punctuation)
words in the reference: 10
words in the MT output: 9

1-gram precision:
9/9 = 100%

1-gram recall:
9/10 = 90%

Are 1-grams enough?

Reference: **It will be considered as** a **sort of bridge .**

MT output: **It will sort of bridge be considered as .**

1-gram precision:
9/9 = 100%

1-gram recall:
9/10 = 90%

rather high values
and what about word order?
⇒ add *n*-grams with higher orders

2-gram precision and recall

Reference: **It~will** will~be **be~considered considered~as** as~a a~sort **sort~of of~bridge** bridge~.

MT output: **It~will** will~sort **sort~of of~bridge** bridge~be **be~considered considered~as** as~.

matches (bold): 4

2-grams in the reference: 9

2-grams in the MT output: 8

2-gram precision:

4/8 = 50.0%

2-gram recall:

4/9 = 44.4%

3-gram precision and recall

Reference: It~will~be will~be~considered **be~considered~as** considered~as~a as~a~sort a~sort~of **sort~of~bridge** of~bridge~.

MT output: It~will~sort will~sort~of **sort~of~bridge** of~bridge~be bridge~be~considered **be~considered~as** considered~as~.

matches (bold): 2

3-grams in the reference: 8

3-grams in the MT output: 7

3-gram precision:

2/7 = 28.6%

3-gram recall:

2/8 =25.0%

higher *n*-gram orders

- we can continue and continue
- however, at some point, precisions and recalls will become 0
- which *n*-gram order to use for evaluation?
  - combine all orders up to Nth

In MT evaluation, usually N=4

– how to combine precisions/recalls of different *n*-gram orders?

joining *n*-gram orders (up to *n*=N)

- arithmetic mean

$$precision = \frac{1\text{-}gram\ precision + 2\text{-}gram\ precision + \ldots + N\text{-}gram\ precision}{N}$$

$$recall = \frac{1\text{-}gram\ recall + 2\text{-}gram\ recall + \ldots + N\text{-}gram\ recall}{N}$$

- geometric mean

$$precision = \sqrt[N]{1\text{-}gram\ precision \cdot 2\text{-}gram\ precision \cdot \ldots \cdot N\text{-}gram\ precision}$$

$$recall = \sqrt[N]{1\text{-}gram\ recall \cdot 2\text{-}gram\ recall \cdot \ldots \cdot N\text{-}gram\ recall}$$

Which averaging is better?

- no evidence that one is better than other
- however, geometric mean has a disadvantage:
  – even if one single *n*-gram value is 0, the overall value becomes 0
  – values for higher order *n*-grams are often 0

What is better: precision or recall?

- both can be used!

$$F_\beta = (1+\beta^2)\frac{2 \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

- recall-oriented metrics correlate better with human adequacy-oriented scores

- BLEU: only uses precision + “brevity penalty” instead of recall in order not to assign high scores to short sentences which do not cover reference translation well

What is better: precision or recall?

- both can be used!

  F-score:

$$F_\beta = (1 + \beta^2) \frac{2 \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

- “beta” > 1 ⇒ more weight on recall
- “beta” < 1 ⇒ more weight on precision
- “beta” = 1 ⇒ precision and recall equally important

Metrics based on n-gram matching

- F-score never used directly for MT evaluation
  (no particular reason)

- the first metric proposed for MT evaluation:

  BLEU (proposed in 2002)

  based on 4-gram precision geometrically averaged “brevity penalty” instead of recall

Metrics based on n-gram matching

- BLEU correlates decently with human scores
- however, a number of better metrics have been developed and proposed
- some of them are used in practice, such as:
  - METEOR (proposed in 2005) based both on precision and recall with more weight on recall
  - chrF (proposed in 2015) F-score based on character n-grams instead of word n-grams
- nevertheless, BLEU is still the most used one

Edit distance: Word Error Rate (WER)

Minimum number of edit operations necessary to transform an MT output into a reference translation

edit operations:

- substitution: replace one word with another
- insertion: add word
- deletion: drop word

Word Error Rate (WER)

Edit (Levenshtein) distance: minimum number of edit operations

WER: normalised Levenshtein distance (divided by the length of the reference translation)

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

higher WER indicates lower MT quality

Word Error Rate (WER)

Reference translation:

Israeli officials are responsible for airport security

System output:

Israeli official responsible airport is security

What is the 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 = 57.1%

WER Calculation

Given a reference translation and an MT system output

how can we calculate the WER score?

WER Calculation

reference translation

MT output

	init.	Israeli	officials	are	responsible	for	airport	security
initialisation	0	1	2	3	4	5	6	7
Israeli	1							
official	2							
responsible	3							
airport	4							
is	5							
security	6							

Create a word alignment matrix between the reference and the MT output with costs for each transition from pair (ref1, out1) to (ref2, out2)

WER Calculation

reference translation

MT output

	init.	Israeli	officials	are	responsible	for	airport	security
initialisation	0	1	2	3	4	5	6	7
Israeli	1	**0**	1	2	3	4	5	6
official	2	1	**1**	**2**	3	4	5	6
responsible	3	2	2	2	**2**	**3**	4	5
airport	4	3	3	3	3	3	**3**	4
is	5	4	4	4	4	4	**4**	4
security	6	5	5	5	5	5	5	**4**

Add the costs for each transition from one cell to another

Insertion	→
Deletion	↓
Substitution	↘
Match	↘

WER Calculation

reference translation

MT output

	init.	Israeli	officials	are	responsible	for	airport	security
	0	1	2	3	4	5	6	7
Israeli	1	0	1	2	3	4	5	6
official	2	1	1	2	3	4	5	6
responsible	3	2	2	2	2	3	4	5
airport	4	3	3	3	3	3	3	4
is	5	4	4	4	4	4	4	4
security	6	5	5	5	5	5	5	4

Find a shortest path from top-left corner to bottom-right corner

Insertion	→
Deletion	↓
Substitution	↘
Match	↘

Problem with WER

Consider a large but correct difference in word order

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

edit distance = 5

WER = 5/7 = 71.4% very high!

Problem with WER

reference translation:

Israeli officials are responsible for airport security

another system output:

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

this translation is good, but with a different word order than the reference translation

⇒ **very high WER score!**

(6 substitutions, 4 insertions => 10/7 = 142.8%)

Problem with WER

- penalises too strongly differences in word order
- does not correlate well with human scores

  (too many acceptable translations are getting high WER scores)

⇒ TER (Translation Edit Rate)

three WER operations + shift operation

Translation Edit Rate (TER)

- as WER, based on number of edit operations required to change an MT output into a reference translation

- additional cost for shifts of words and word sequences

$$TER = \frac{insertions + deletions + substitutions + shifts}{reference\ length}$$

Translation Edit Rate (TER)

Reference: It will be considered as a sort of bridge .

MT output: It will sort of bridge be considered as .

WER:

Reference: It will **be**$_{del.}$ **considered**$_{del.}$ **as**$_{del}$ **a**$_{del.}$ sort of bridge .

MT output: It will sort of bridge **be**$_{ins.}$ **considered**$_{ins.}$ **as**$_{ins.}$.

edit distance = 4 deletions + 3 insertions = 7

WER = 7/10

TER:

Reference: It will **[be considered as]**$_{shift}$ **a**$_{del.}$ sort of bridge .

MT output: It will sort of bridge **[be considered as]**$_{shift}$.

Translation Edit Rate (TER)

Reference: It will be considered as a sort of bridge .

MT output: It will sort of bridge be considered as .

WER:

Reference: It will **be**$_{del.}$ **considered**$_{del.}$ **as**$_{del}$ **a**$_{del.}$ sort of bridge .

MT output: It will sort of bridge **be**$_{ins.}$ **considered**$_{ins.}$ **as**$_{ins.}$.

TER:

Reference: It will **[be considered as]**$_{shift}$ **a**$_{del.}$ sort of bridge .

MT output: It will sort of bridge **[be considered as]**$_{shif}$.

edit distance = 1 deletion + 1 shift = 2

TER = 2/10

Translation Edit Rate (TER)

- the most widely used metric based on edit distance
- proposed in 2006 (four years after BLEU)

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

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

Error Classification

identify and classify errors in a translated text

Reference translation:

Israeli **officials are** responsible **for** airport security

System output:

Israeli **official** responsible airport **is** security

"official" = incorrect word form (number of a noun)

"is" = incorrect word form (verb person) and
incorrect word order (position)

"for" = missing word (preposition)

Why Error Classification?

- identify type and nature of translation errors – both MT and human translations
- useful for adjusting MT systems
- useful for reporting back to clients
- LSPs use it to monitor translators' work

Human Error Classification

- more effort than assigning scores
  – thinking about the type of the error (which class it belongs to)
  – a number of errors could belong to more than one class (ambiguous)

- subjective (often more ways to interpret an error)
  – IAA generally lower than for overall scores

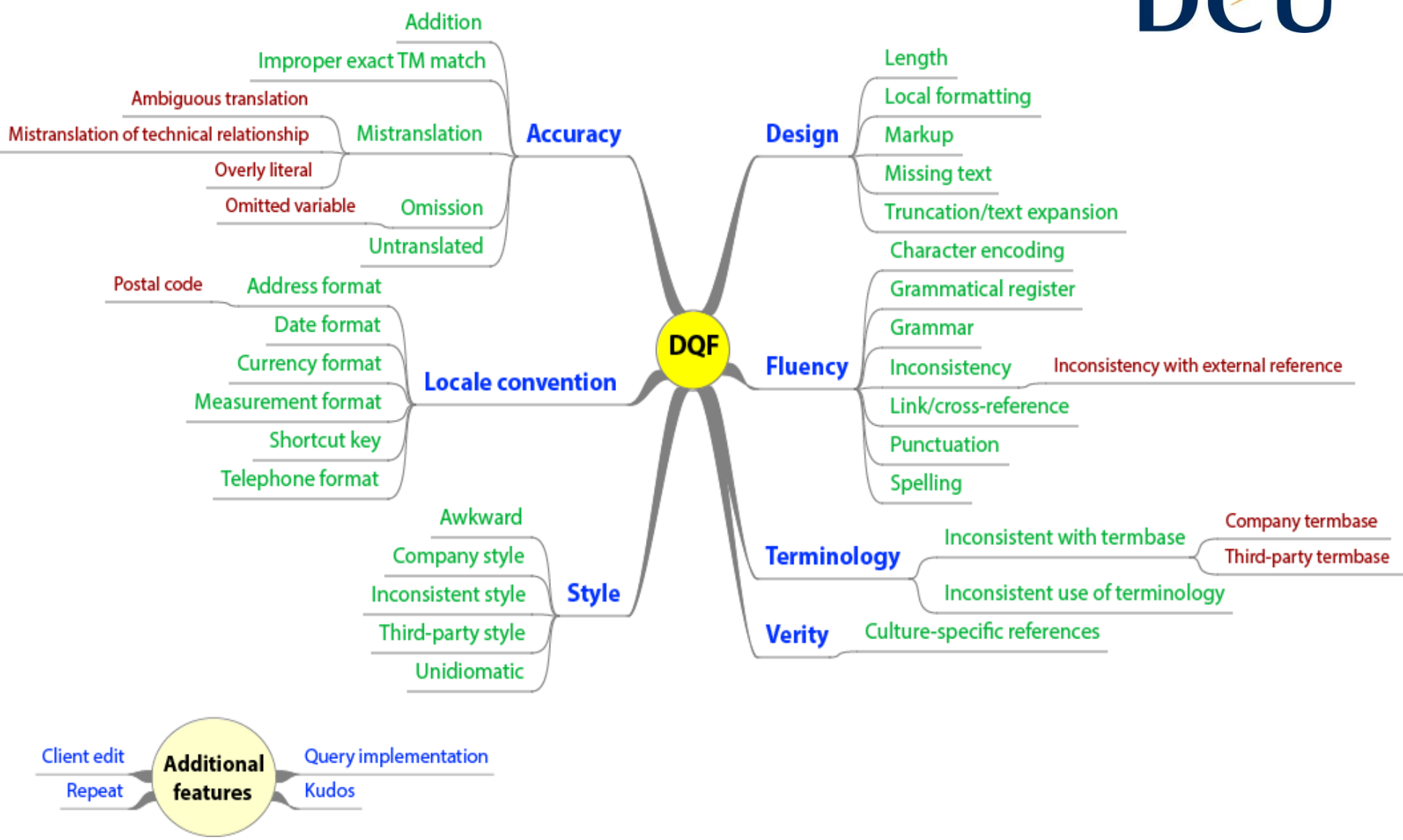
- even organisation is not trivial
  – choose/define a meaningful scheme (typology)
  – prepare guidelines for evaluators

Error Typologies

- from very simple to very detailed/fine grained
- more fine-grained provides more information but
  - more difficult for annotators
  - lower inter-annotator agreements
  - require more elaborated guidelines

- how to define error typology?
  - depends on the task and the goal
  - there are existing typologies to be adapted

Multidimensional Quality Metric (MQM)



Multidimensional Quality Metric (MQM)

- can be used both for MT and HT evaluation
- not necessary to use the full scheme
- a sub-set of error types is usually selected according to the task at hand

Automatic Error Classification

- motivation: similar to automatic evaluation metrics – to reduce the time and effort
- much less work than on automatic evaluation metrics
- so far only coarse-level typologies supported (e.g. distinguishing morphological, syntactic and lexical errors)

Quality Estimation

- automatic estimation of translation quality from the source text and its translation

- no comparison against a reference translation (unlike automatic evaluation and automatic error classification)

Quality Estimation

no reference translation

- no bias towards one possible correct translation
- reference translations usually not available in the "real world"

on the other hand:

- requires labelled training data and model

  (similarly to MT itself, but generally simpler models and less training data)

Quality Estimation

- definition of quality
- labelled data
- model

Quality Estimation: Quality

What to estimate?

- adequacy?
- fluency?
- both together?
- comprehension?
- effort required to post-edit (correct) the output?
- an automatic evaluation score?
- binary error labels for each word (“error”/”correct”)?
- error labels according to an error scheme?
- …

There is a merit in each of these possibilities!

Quality Estimation: Data

What type of data is needed?

labelled data:

- source text
- its MT output
- quality labels
  – on the sentence level
  – on the word level

Quality Estimation: Model

What is the “model” which can predict quality labels?

a classifier able to predict the quality labels

- before: different machine learning algorithms based on various text characteristics (features)
- nowadays: based on neural networks
  – network input: source text and MT output
  – network output: quality labels

Exercise

- **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 .*

Calculate 3-gram recall, 3-gram precision, WER and TER for each translation.

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.