MT Evaluation Basics

April 17, 2023

Koel Dutta Chowdhury

koeldc@lst.uni-saarland.de

Machine Translation

Summer 2023

Slides: Josef van Genabith

DFKI GmH

Josef.van_Genabith@dfki.de

Evaluating MT

UNIVERSITÄT DES SAARLANDES

- What do we want to know?
  - How good is the MT output?
  - Is a system useful?
  - Is one system better than another?
  - What errors does it make?

Evaluating MT

- Why do we want to evaluate MT? E.g. to help it get better …:

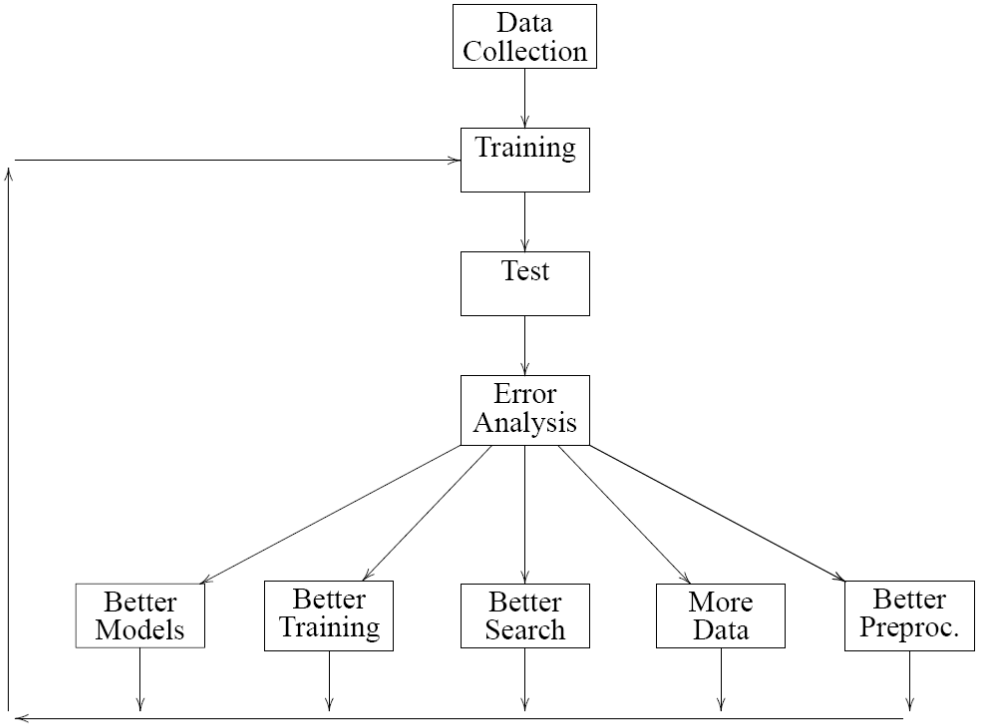


Figure 3.1: Development cycle of a statistical MT system.

(Och 2000)

Evaluating MT

UNIVERSITÄT DES SAARLANDES

- What do we want to know?
  - How good is the MT output?
  - Is a system useful?
  - Is one system better than another?
  - What mistakes does it make?

- When is a translation a good translation?
  - Equivalent in meaning to source text: Adequacy
  - Fluent in target language: Fluency

Evaluating MT

UNIVERSITÄT DES SAARLANDES

- What do we want to know?
  - How good is the MT output?
  - Is a system useful?
  - Is one system better than another?
  - What mistakes does it make?

- When is a translation a good translation?
  - Equivalent in meaning to source text: Adequacy
  - Fluent in target language: Fluency

- What do we compare with? How many good translations are there?
- Is there a single “best” translation?

Ten Translations of a Chinese Sentence

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.

(a typical example from the 2001 NIST evaluation set)

Evaluating MT

UNIVERSITÄT DES SAARLANDES

- How do we evaluate MT?
  - Manual (“human”, “subjective”)
  - Automatic (“objective”)
  - Quality (overall, a number like BLEU, TER, human score etc.)
  - Diagnostic (what goes wrong, linguistic analysis)
  - Scoring
  - Ranking
  - Intrinsic
  - Extrinsic

Evaluating MT

UNIVERSITÄT DES SAARLANDES

- How do we evaluate MT?
  - Manual (“human”, “subjective”)
  - Automatic (“objective”)

- Manual/human/”subjective” (+ intrinsic)
  - Human professional translators
  - People proficient in source and target language at stake
  - People who only understand target but have access to a reference
  - The Crowd … turkers …

  - Often very time consuming and expensive
  - Not easy to reproduce: rater/inter-annotator agreement
  - MT output hard to rate … (sometimes subtle, sometimes bad …)
  - Still: the yardstick, the gold-standard, the ideal …

Human Evaluation

UNIVERSITÄT DES SAARLANDES

- How good/bad, general quality, scoring
  - rating on a scale or absolute rating = assigning a score to a translation
  - comparing two translations and deciding which of the two is better or worse
- What is wrong: diagnostic evaluation …

Human Evaluation: Scoring

UNIVERSITÄT DES SAARLANDES

- Guidelines

- Adequacy (Scale of 5) (~ “meaning”)
  - 5 = All meaning
  - 4 = Most meaning
  - 3 = Much meaning
  - 2 = Little meaning
  - 1 = none
- Fluency (Scale of 5) (~ “grammar”)
  - 5 = Flawless (English)
  - 4 = Good (English)
  - 3 = Non-native (English)
  - 2 = Disfluent (English)
  - 1 = Incomprehensible

Human Evaluation: Scoring

UNIVERSITÄT DES SAARLANDES

- Guidelines

- Adequacy (Scale of 5) (~ “meaning”)
  - 5 = All meaning
  - 4 = Most meaning
  - 3 = Much meaning
  - 2 = Little meaning
  - 1 = none
- Fluency (Scale of 5) (~ “grammar”)
  - 5 = Flawless (English)
  - 4 = Good (English)
  - 3 = Non-native (English)
  - 2 = Disfluent (English)
  - 1 = Incomprehensible

Very hard to operationalise!

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 .

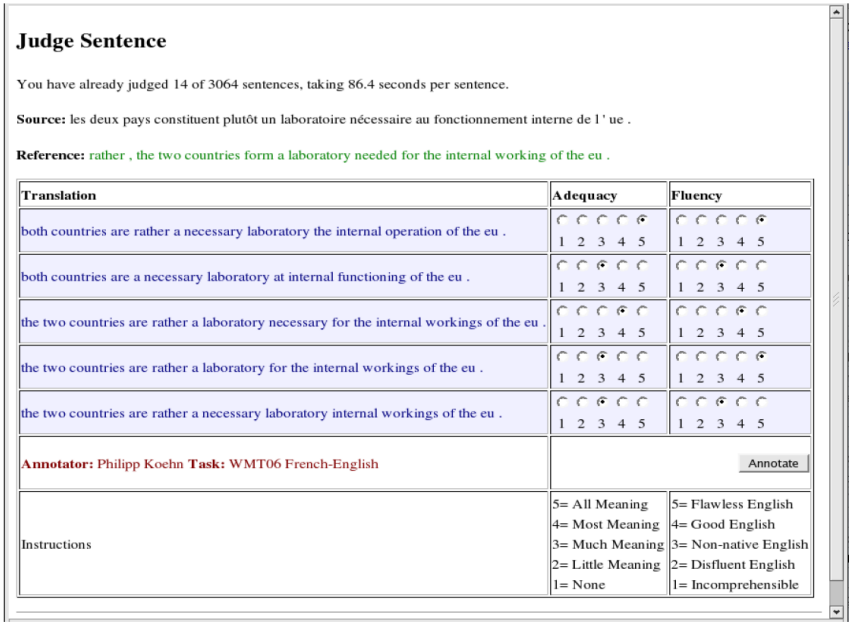
Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings 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: Scoring

UNIVERSITÄT DES SAARLANDES

- Very hard to do for humans
- Juggle 5 (possibly equally miserable or good) automatic translations (for possibly long sentences)
- With respect to 2 dimensions on a scale of 5 each …
- Poor inter-annotator/rater agreement!
- Many data points (raters) + statistics = good results possible …

- Don't know yet what is wrong
- why a system is good or bad
- …



Judge Sentence

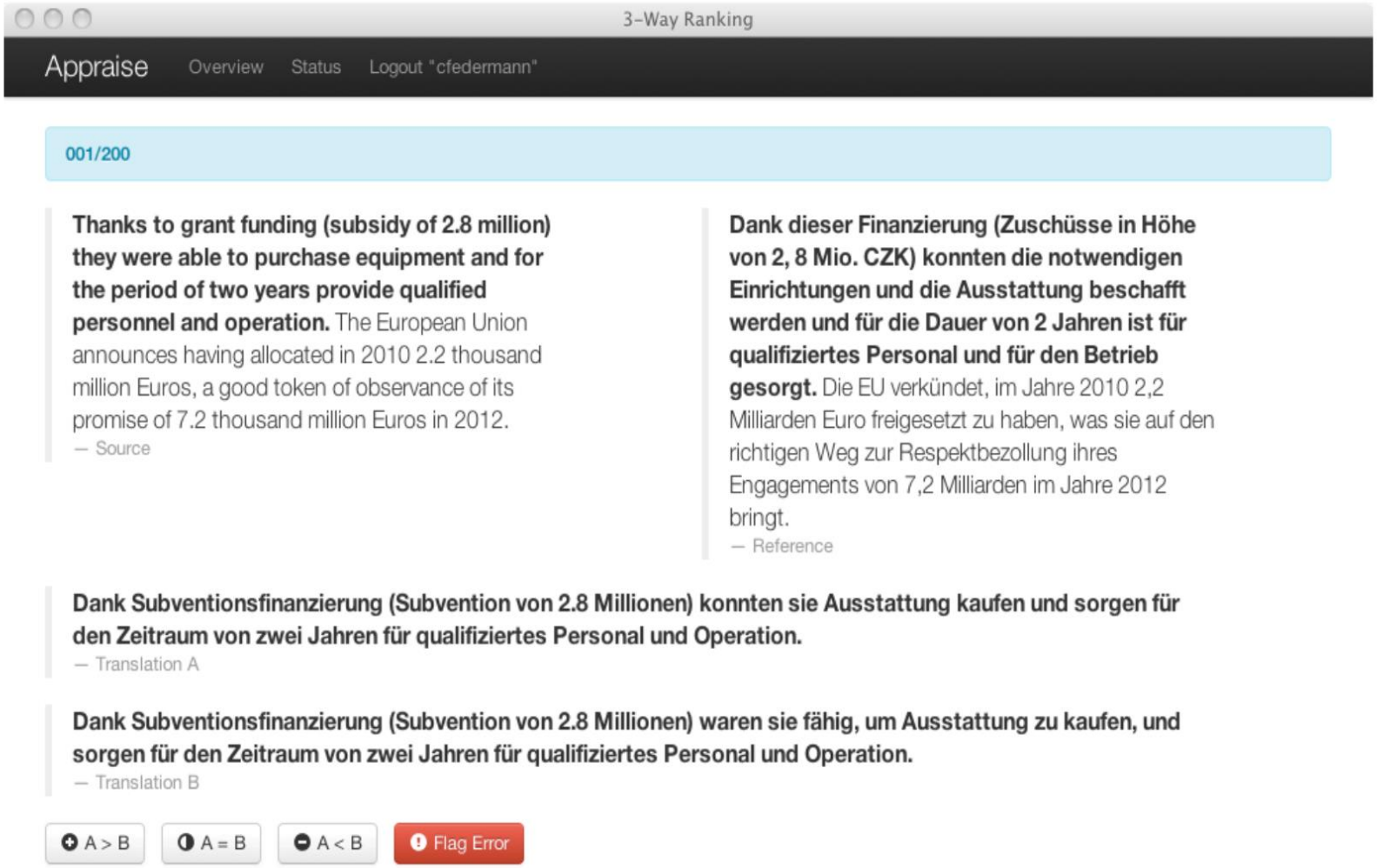
You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

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: Ranking



Human Evaluation: Diagnostic

- What is wrong/diagnostic evaluation: error classifications (Vilar et al. 2006):

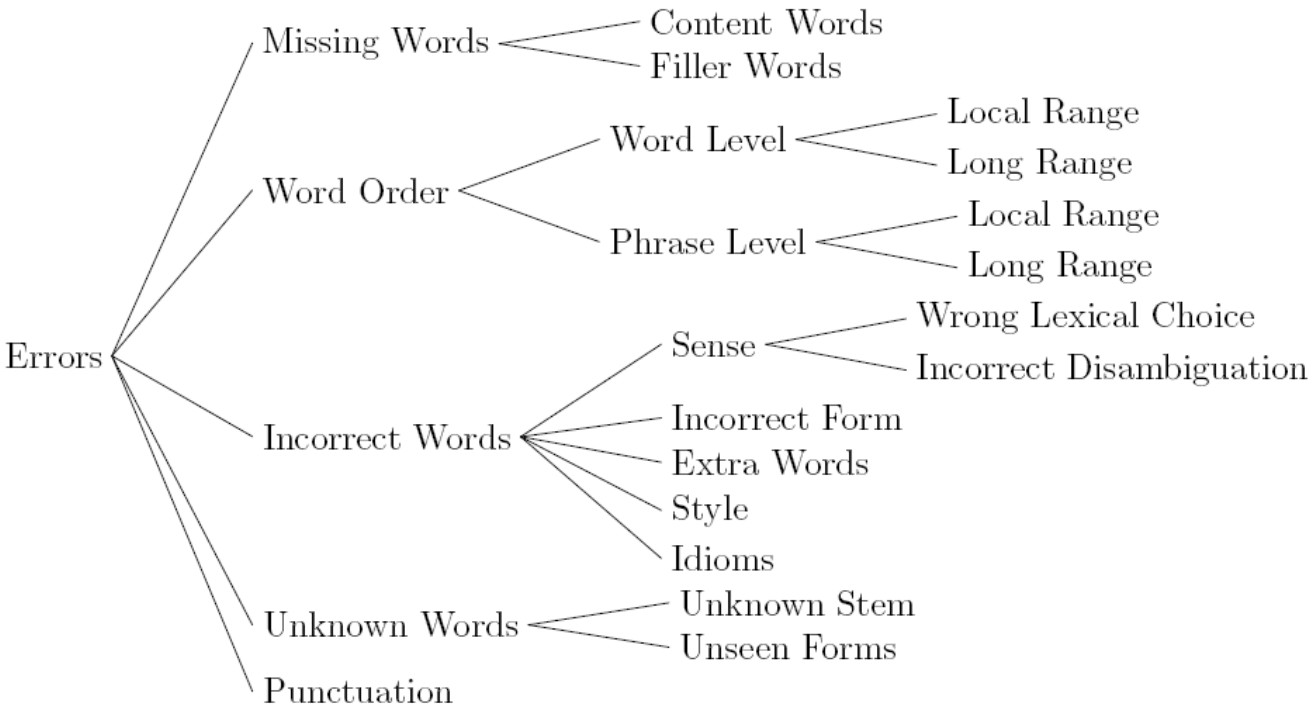
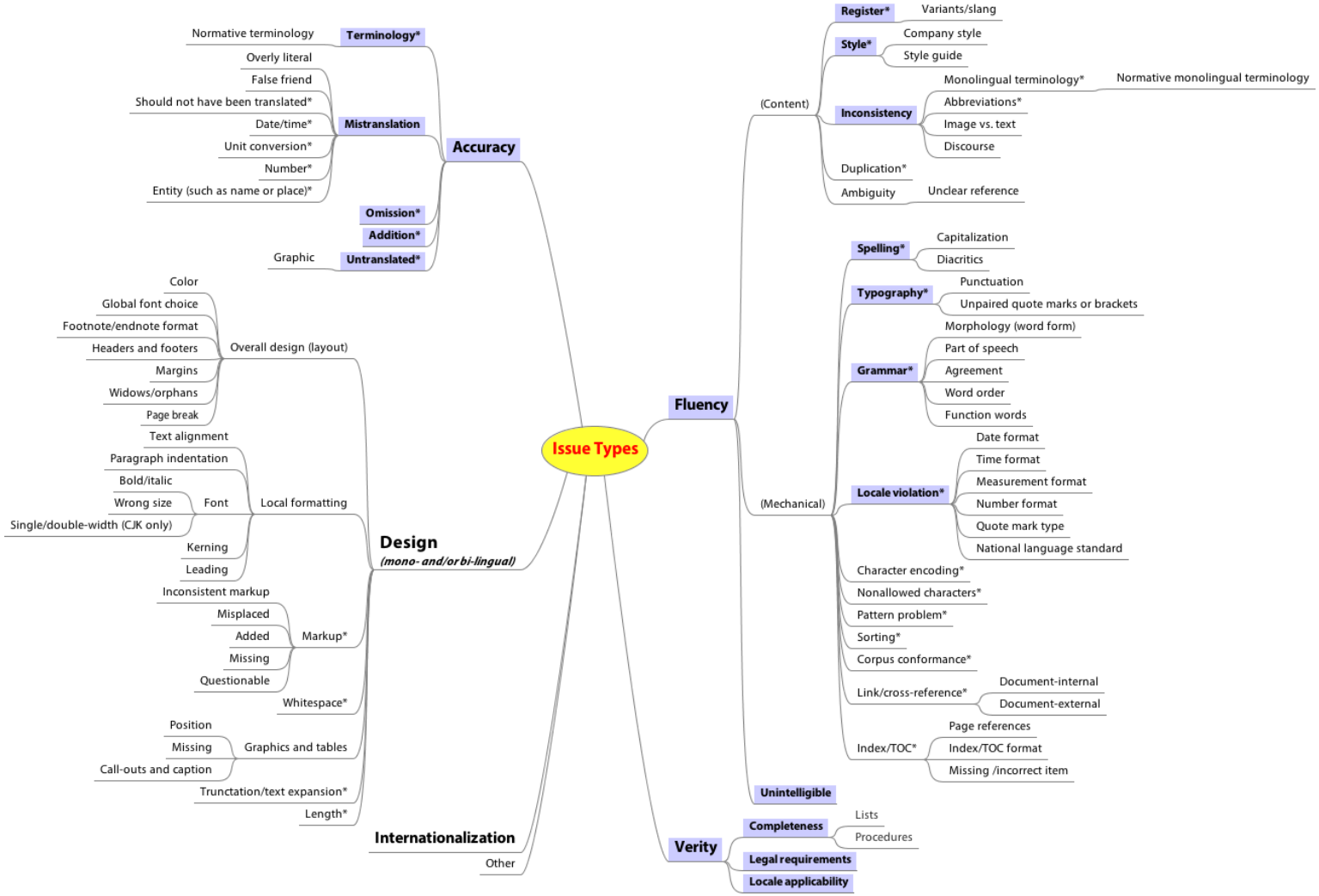


Figure 1: Classification of translation errors.

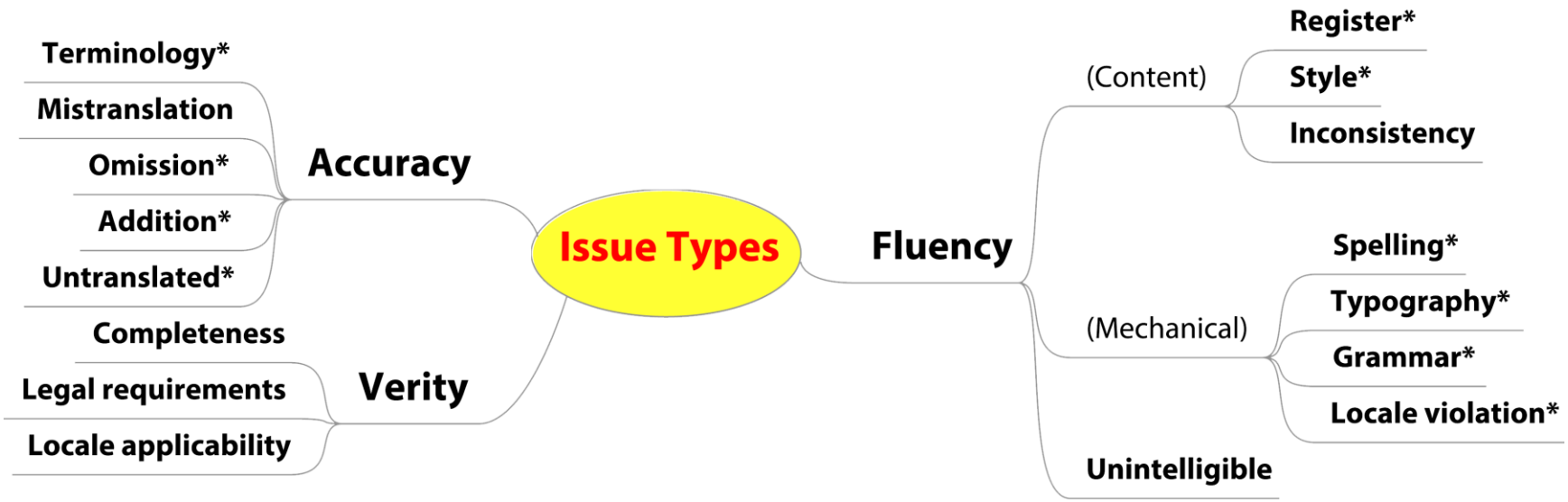
Human Evaluation: Diagnostic

- Error classification MQM QTLaunchPad (Lommel et al. 2013):



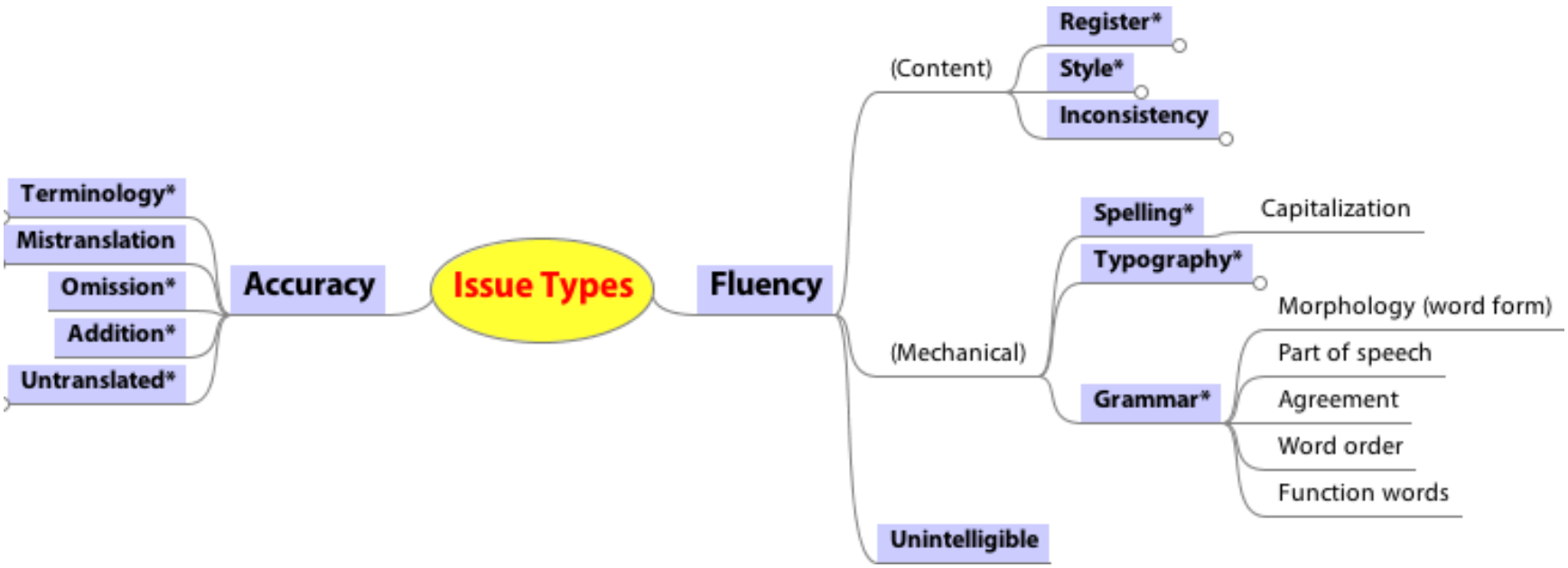
Human Evaluation: Diagnostic

- Error classification MQM Core QTLaunchPad (Lommel et al. 2013):



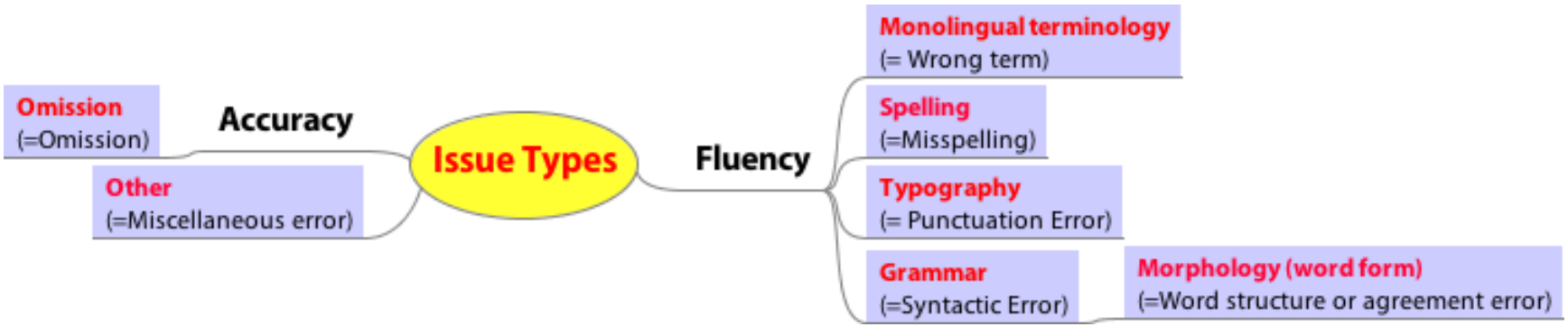
Human Evaluation: Diagnostic

- Error classification MQM MT Subset (Lommel et al. 2013):



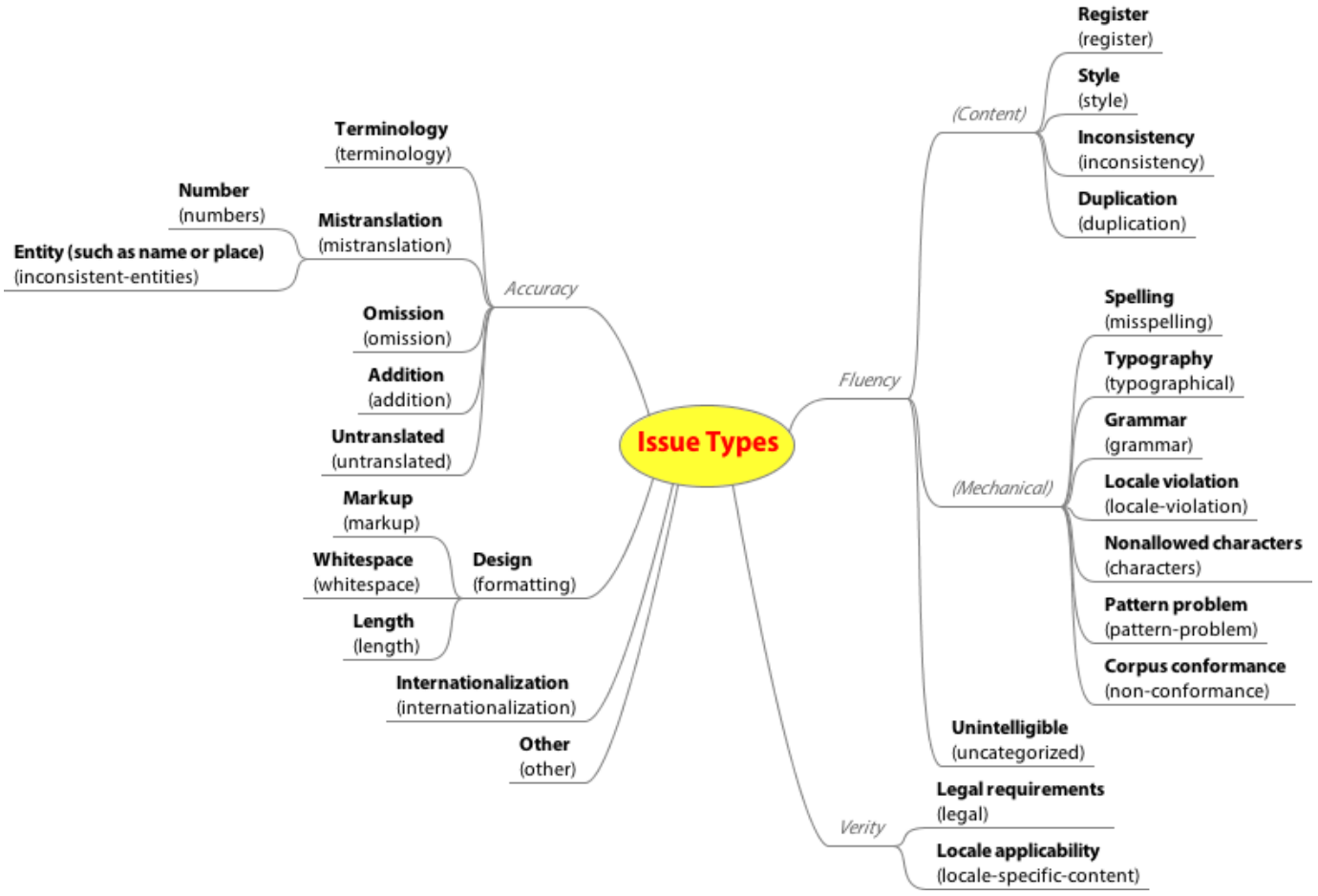
Human Evaluation: Diagnostic

- Error classification MQM mapping to SAE J2450 (Lommel et al. 2013):



Human Evaluation: Diagnostic

- Error classification MQM mapping to ITS 2.0 (Lommel et al. 2013):



Human Evaluation

UNIVERSITÄT DES SAARLANDES

We have seen quite a few variants

- Fluency/Adequacy (with score on Likert scale): scoring
  - Complex, high granularity, inter- and intra-rater agreement, time consuming
- Simple ranking: better, worse, same
  - Easier, faster, but still very time consuming

- Many data points and statistics needed: “direct assessment”

Don’t know what’s wrong! Diagnostic evaluation:

- Error typology/annotation: Vilar et al. 2006, MQM, DQF, ..
  - Time consuming, expensive, low rater agreement …

Human Evaluation

UNIVERSITÄT DES SAARLANDES

We have seen quite a few variants

- Fluency/Adequacy (with score on Likert scale)
- Simple ranking: better, worse, same … and from the pairwise rankings we can computer overall rankings (= scores)
- Error annotation: Vilar et al. 2006, MQM, DQF, ..

- Can you think about another way/dimension?

Human Evaluation

UNIVERSITÄT DES SAARLANDES

We have seen quite a few variants

- Fluency/Adequacy (with score on Likert scale)
- Simple ranking: better, worse, same … and from the pairwise rankings we can computer overall rankings (= scores)
- Error annotation: Vilar et al. 2006, MQM, DQF, ..

- Can you think about another way/dimension?

- Intrinsic vs. extrinsic evaluation (extrinsic = task-based)

Human Evaluation: Extrinsic

UNIVERSITÄT DES SAARLANDES

- Intrinsic vs. extrinsic evaluation (task-based)
- Extrinsic evaluation: “external” task
- Given two MT systems A and B, the better one is the one that e.g. requires less human post-editing (correction) to eliminate mistakes and produce a good translation …
- Extrinsic Task: translation and post-editing
- Better = less time = less effort = less cost …
- Other possible extrinsic tasks: e.g. answer questions from the translation, … many others possible

Human Evaluation:

- Scoring – ranking, diagnostic, …, intrinsic – extrinsic etc.
- Time consuming
- Expensive
- Difficult to define and operationalise
- Hard to reproduce: low inter- and intra-rater agreement
- Hard to scale: though see crowd-sourcing (Chris Callison-Burch papers, Yvette Graham et al. papers on Direct Assessment)

- Still: human eval. indispensable and the yardstick
- All “serious” MT shared tasks/competitions (such as WMT, IWSLT, NIST, …) do a human evaluation track

- and, of course, they also do automatic evaluation …

Automatic Evaluation

UNIVERSITÄT DES SAARLANDES

- The basic idea:
- Given a human professional reference translation (or several reference translations), compare MT output against this
- How?
- How similar are they?

- Word-, n-gram-, character-, string-overlap (surface similarity …)
- F-measure, BLEU, …, chrF3, …
- More sophisticated stuff (not just surface string matching based)
  - Stemming, morphological analysis, synonyms, paraphrases, syntactic and semantic structure, etc. METEOR
- There are also edit-distance based metrics: TER, characTER, …
- Neural methods: ReVaL, …

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: Israeli officials responsibility of airport safety

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: Israeli officials responsibility of airport safety

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety

- Word overlap: precision, recall and f-measure
- Precision: how many words in MT output are correct (in ref)?
- Recall: how many of the words in reference are in the MT output?
- F-measure: harmonic mean of precision and recall

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security

System A: Israeli officials responsibility of airport safety

- Word overlap: precision, recall and f-measure
- Precision: how many of the words in output are correct (in ref)?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{3}{6} = 0.5$$

- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{3}{7} = 0.43$$

- F-measure: harmonic mean of precision and recall

$$f\_score = \frac{2 \times precision \times recall}{precision + recall} = 0.46$$

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli

- Why not precision alone?

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security
System A: Israeli

- Why not precision alone?
- Precision: how many of the words in output are correct (in ref)?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{1}{1} = 1$$

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security
System A: Israeli

- Why not precision alone?

- Precision: how many of the words in output are correct (in ref)?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{1}{1} = 1$$

- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{1}{7}$$

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security
System A: Israeli

- Why not precision alone?

- Precision: how many of the words in output are correct (in ref)?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{1}{1} = 1$$

- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{1}{7}$$

- F-score: harmonic mean of precision and recall

$$f\_score = \frac{2 \times precision\ \times recall}{precision + recall} = 0.25$$

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: “dump of all the words in MT vocabulary/lexicon …”

- Why not recall alone?

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security

System A: “dump of all the words in MT vocabulary/lexicon …”

- Why not recall alone?
- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{7}{7} = 1$$

Automatic Evaluation: F-Measure

Reference: Israeli officials are responsible for airport security

System A: "dump of all the words in its vocabulary/lexicon …"

- Why not recall alone?
- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{7}{7} = 1$$

- Precision: how many of the words in output are correct?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{7}{50{,}000}$$

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: "dump of all the words in its vocabulary/lexicon …"

- Why not recall alone?

- Recall: how many of the words in reference are in the output?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ reference} = \frac{7}{7} = 1$$

- Precision: how many of the words in output are correct?

$$\frac{\#\ correct\ words\ in\ output}{\#\ total\ words\ in\ output} = \frac{7}{50{,}000}$$

- F-score: harmonic mean of precision and recall

$$f\_score = \frac{2 \times precision \times recall}{precision + recall} \approx 0.00006$$

Precision, Recall, F-Score, Accuracy …

		Predictions: true	Predictions: false
Ground Truth	true	tp	fn
	false	fp	tn

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fn}$$

$$F = \frac{2 \times P \times R}{P + R}$$

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}$$

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference:	Israeli officials are responsible for airport security
System A:	Israeli officials responsibility of airport safety
System B:	airport security Israeli officials are responsible
System C:	security Israeli are officials responsible airport

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: Israeli officials responsibility of airport safety

System B: airport security Israeli officials are responsible

System C: security Israeli are officials responsible airport

	System A	System B	System C
precision	0.50	1.00	1.00
recall	0.43	0.86	0.86
f-score	0.46	0.92	0.92

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety
System B: airport security Israeli officials are responsible
System C: security Israeli are officials responsible airport

	System A	System B	System C
precision	0.50	1.00	1.00
recall	0.43	0.86	0.86
f-score	0.46	0.92	0.92

- Problem: f-measure can reward unintelligible word salad in C if individual words are O.K. …
- Fails to reflect …. What?

Automatic Evaluation: F-Measure

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety
System B: airport security Israeli officials are responsible
System C: security Israeli are officials responsible airport

	System A	System B	System C
precision	0.50	1.00	1.00
recall	0.43	0.86	0.86
f-score	0.46	0.92	0.92

- Problem: f-measure can reward unintelligible word salad in C if individual words are O.K. …
- Fails to reflect …. word order !!!

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety
System B: airport security Israeli officials are responsible
System C: security Israeli are officials responsible airport

- Look at n-gram overlap, not just word overlap

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety
System B: airport security Israeli officials are responsible
System C: security Israeli are officials responsible airport

- Look at n-gram overlap, not just word overlap
- n-gram precision $(n = 1 \ldots 4)$ times a brevity penalty ("recall")

$$BLEU := \min\left(1, \exp\left(1 - \frac{|reference|}{|output|}\right)\right)\left(\prod_{n=1}^{4} n - gram\ precision\right)^{\frac{1}{4}}$$

- $BLEU = 0$ if the hypothesis does not have at least one matching n-gram for any one of the $n = 1 \ldots 4$: systems A and C!

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security

System A: Israeli officials responsibility of airport safety

System B: airport security Israeli officials are responsible

System C: security Israeli are officials responsible airport

$$BLEU := \min(1, \exp(1 - \frac{|reference|}{|output|}))(\prod_{n=1}^{4} n - gram\ precision)^{\frac{1}{4}}$$

$$(\prod_{n=1}^{4} n - gram\ prec)^{\frac{1}{4}} = (\frac{6}{6} \times \frac{4}{5} \times \frac{2}{4} \times \frac{1}{3})^{\frac{1}{4}} = 0.1333^{\frac{1}{4}} = 0.60$$

$$\min\left(1, \exp\left(1 - \frac{|reference|}{|output|}\right)\right) = \min\left(1, \exp\left(1 - \frac{7}{6}\right)\right) = 0.87$$

$$BLEU_B = 0.87 \times 0.60 = 0.52$$

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Reference: Israeli officials are responsible for airport security
System A: Israeli officials responsibility of airport safety
System B: airport security Israeli officials are responsible
System C: security Israeli are officials responsible airport

	System A	System B	System C
f-score	0.46	0.92	0.92
BLEU	0	0.52	0

- Problem: BLEU assigns 0 to many hypotheses …
- Meant to work on document, not individual sentence, level
- sBLEU for sentence level … (smoothed BLEU)

Automatic Evaluation: BLEU

$$BLEU := \min(1, \exp(1 - \frac{|reference|}{|output|}))(\prod_{n=1}^{4} n - gram\ precision)^{\frac{1}{4}}$$

- Fancy way of writing BLEU:

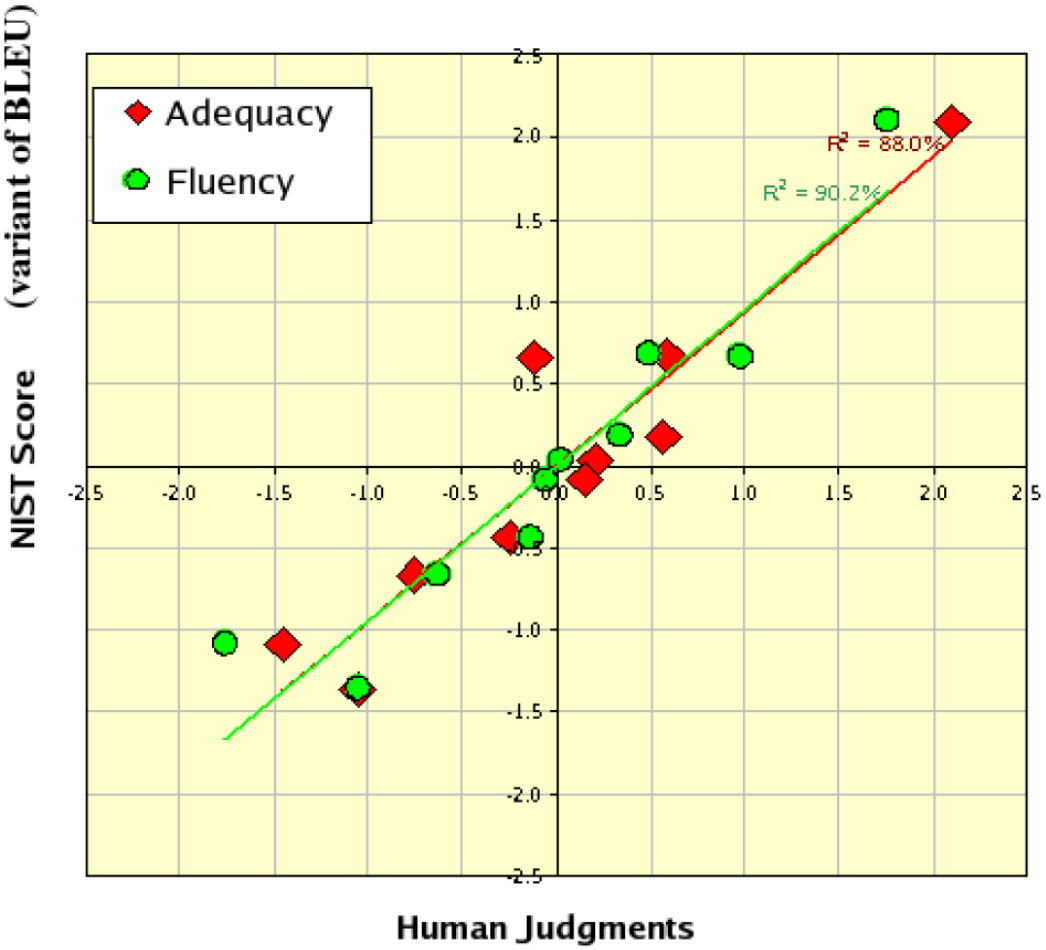
$$BLEU := \min(1, \exp(1 - \frac{|reference|}{|output|}))(\exp\left(\sum_{n=1}^{4} \lambda_n \times \log(n - gram\ prec)\right))^{\frac{1}{4}}$$

- Taking log of n-gram prec, summing over them and using inverse function of log: e …
- λ_i usually 1…

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Correlation with Human Judgement

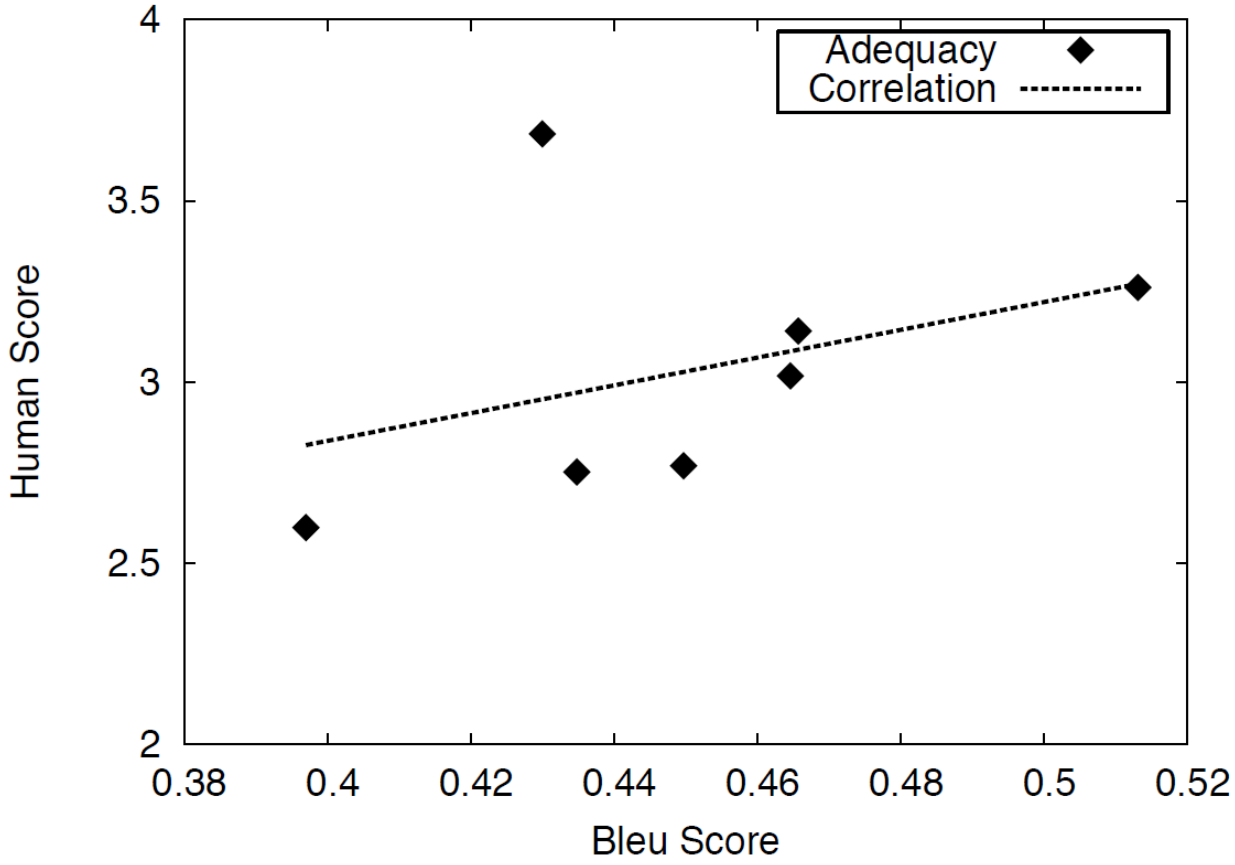


Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Evidence of Shortcomings of Automatic Metrics

Post-edited output vs. statistical systems (NIST 2005)

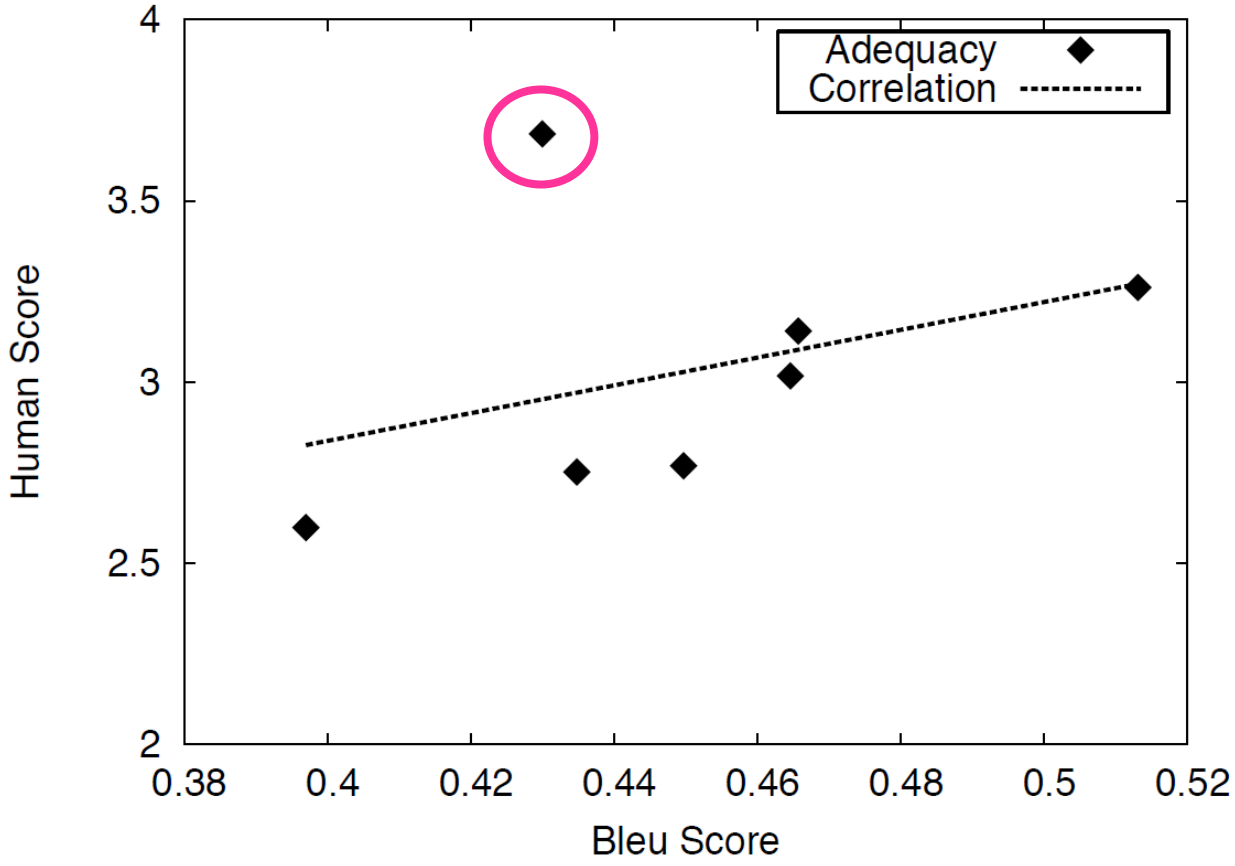


Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Evidence of Shortcomings of Automatic Metrics

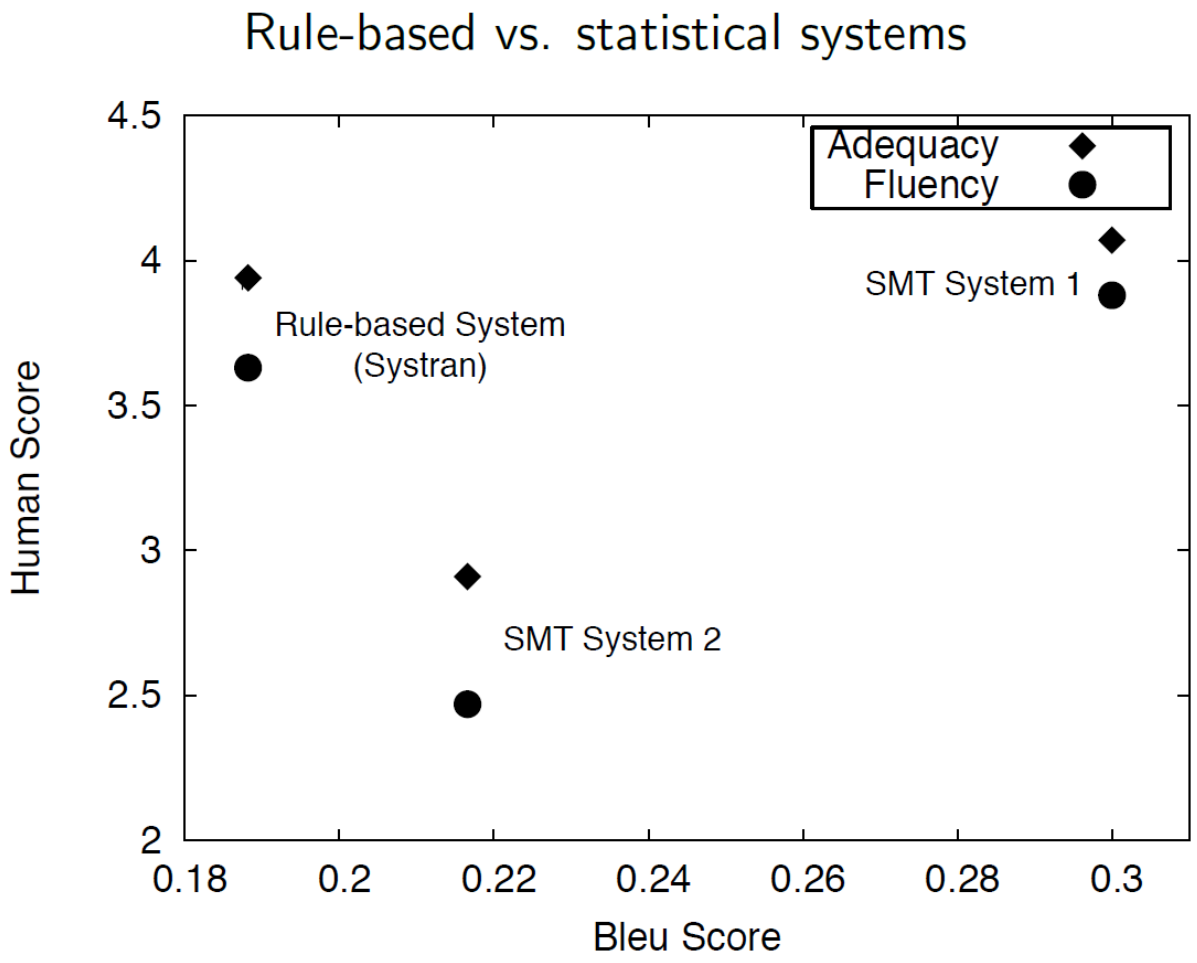
Post-edited output vs. statistical systems (NIST 2005)



Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

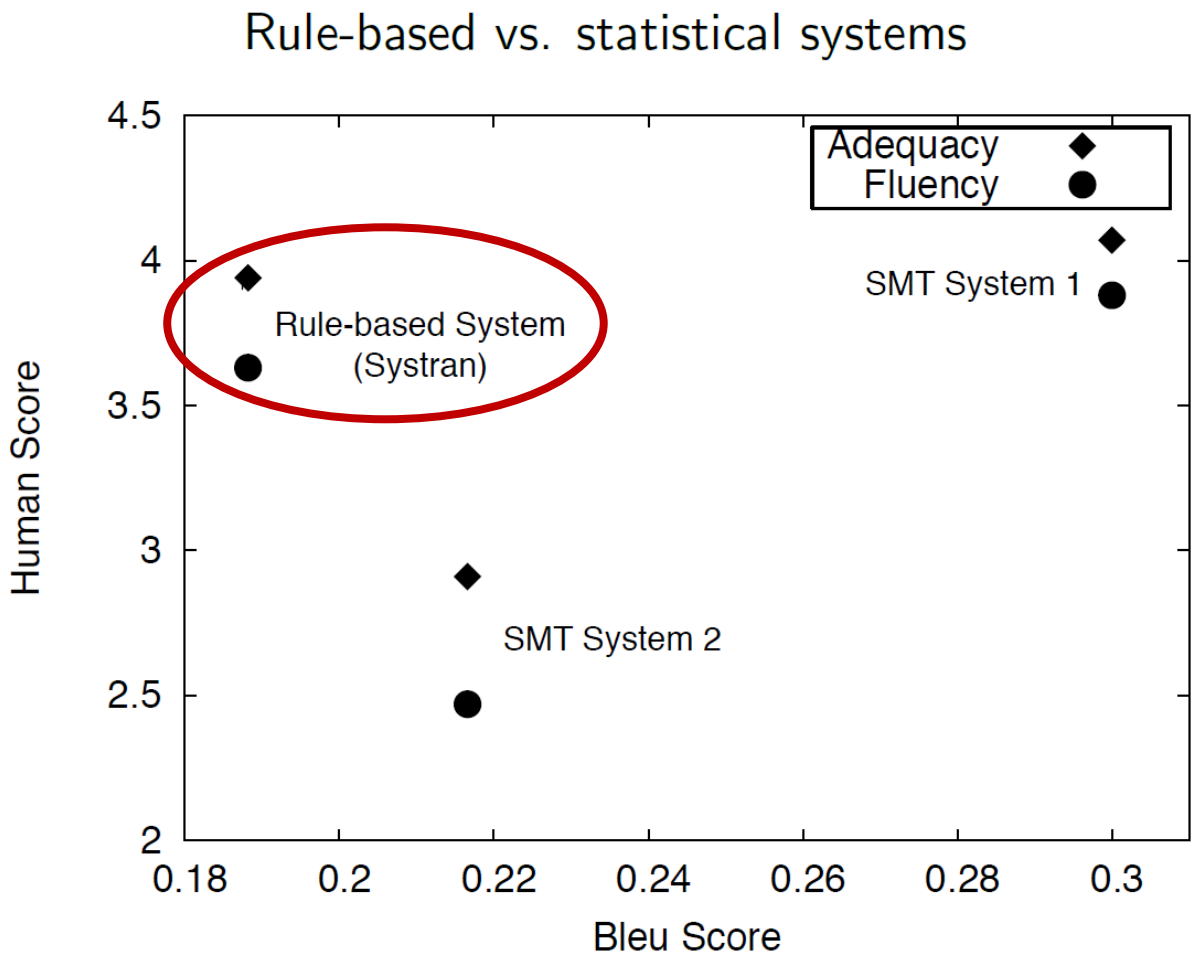
Evidence of Shortcomings of Automatic Metrics



Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

Evidence of Shortcomings of Automatic Metrics



String-based Automatic MT Evaluation Metrics

UNIVERSITÄT DES SAARLANDES

- Consistent
- Treat all words as strings: no difference between function and content words, morphological variants …
- Do not consider global grammaticality
- Do not consider meaning

Yesterday John resigned from the company
John quit the company yesterday

- Punishes perfect paraphrases
- Scores by themselves do not mean much
- Don't tell you what is wrong …
- Human translators often score fairly low on BLEU

Automatic MT Evaluation Metrics

UNIVERSITÄT DES SAARLANDES

- But: we could in principle use many references … expensive again …

- BLEU (n-gram), METEOR (stemming, paraphrases), TER (edit-distance based), dependency-based MT evaluation measures (Karolina Owczarzak), MEANT, chrF3 character- rather than word-token based measures (Maja Poppovich), BEER (trainable), neural MT evaluation measures like ReVal (Rohit Gupta et al.), based on recursive neural networks, … a very active area of research: YiSi-1/2, BERTscore

- Quality Estimation (QE, as opposed to MT evaluation): estimate/predict quality (BLEU/TER scores, post-editing effort) without reference … !!! (papers by Lucia Specia a good starting point): Lucia Specia, Carolina Scarton, Gustavo Paetzold. “Quality Estimation for Machine Translation”. Synthesis Lectures on Human Language Technologies September 2018

Human Evaluation:

UNIVERSITÄT DES SAARLANDES

- Time Consuming
- Expensive
- Difficult to define and operationalise
- Hard to reproduce: inter-rater agreement
- Hard to scale: though see crowd-sourcing (Chris Callison-Burch papers) and (Yvette Graham's papers on Direct Assessment, which is used pretty much as standard for human crowd eval)
- Still: indispensable and the yardstick
- All "serious" MT shared tasks/competitions (such as WMT, IWSLT, NIST, …) do a human evaluation track
- and, of course, they also do automatic evaluation: BLEU, TER, Meteor (chrF3, character, Beer, …)
- WMT Metrics Matter shared task

Some References to Start Out in MT Evaluation

- Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005*
- Christian Federmann. Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. PBML 2012(98)
- Graham, Y. and T. Baldwin. (2014) "Testing for Significance of Increased Correlation with Human Judgment". *Proceedings of EMNLP 2014, Doha, Qatar*
- Rohit Gupta, Constantin Orasan, Josef van Genabith. 2015. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal
- Lommel, Burchard and Uzskoreit. Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality.
- Karolina Owczarzak, Josef van Genabith, Andy Way. Evaluating machine translation with LFG dependencies. Machine Translation volume 21, pages95–119(2007)
- Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318
- Maja Popovic. CHRF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisboa, Portugal,
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.
- Milos Stanojevic and Khalil Sima'an. BEER: BEtter Evaluation as Ranking. Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 414–419, Baltimore, Maryland USA
- David Vilar, Jia Xu, Luis Fernando D'Haro, Hermann Ney. Error Analysis of Statistical Machine Translation Output. LREC 2006
- Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, Hermann Ney. CharacTER: Translation Edit Rate on Character Level. 2016. Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 505–510, Berlin, Germany


TIME FOR QUESTIONS
12
1
2
3
4

Precision, Recall, F-Score, Accuracy …

		Predictions: true	Predictions: false
Ground Truth	true	tp	fn
	false	fd	tn