Statistical Machine Translation

June 3rd, 2014

UNIVERSITÄT DES SAARLANDES

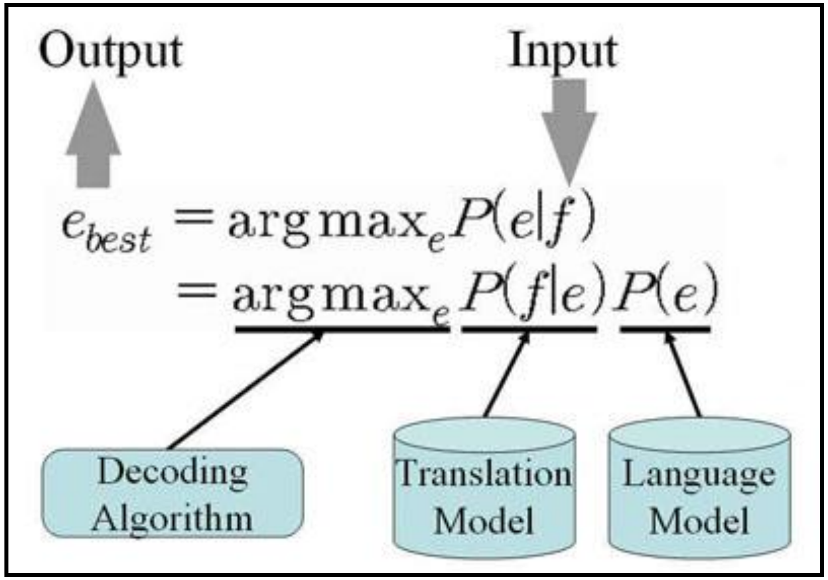
Josef van Genabith
DFKI GmbH
Josef.van_Genabith@dfki.de

Language Technology II
SS 2014

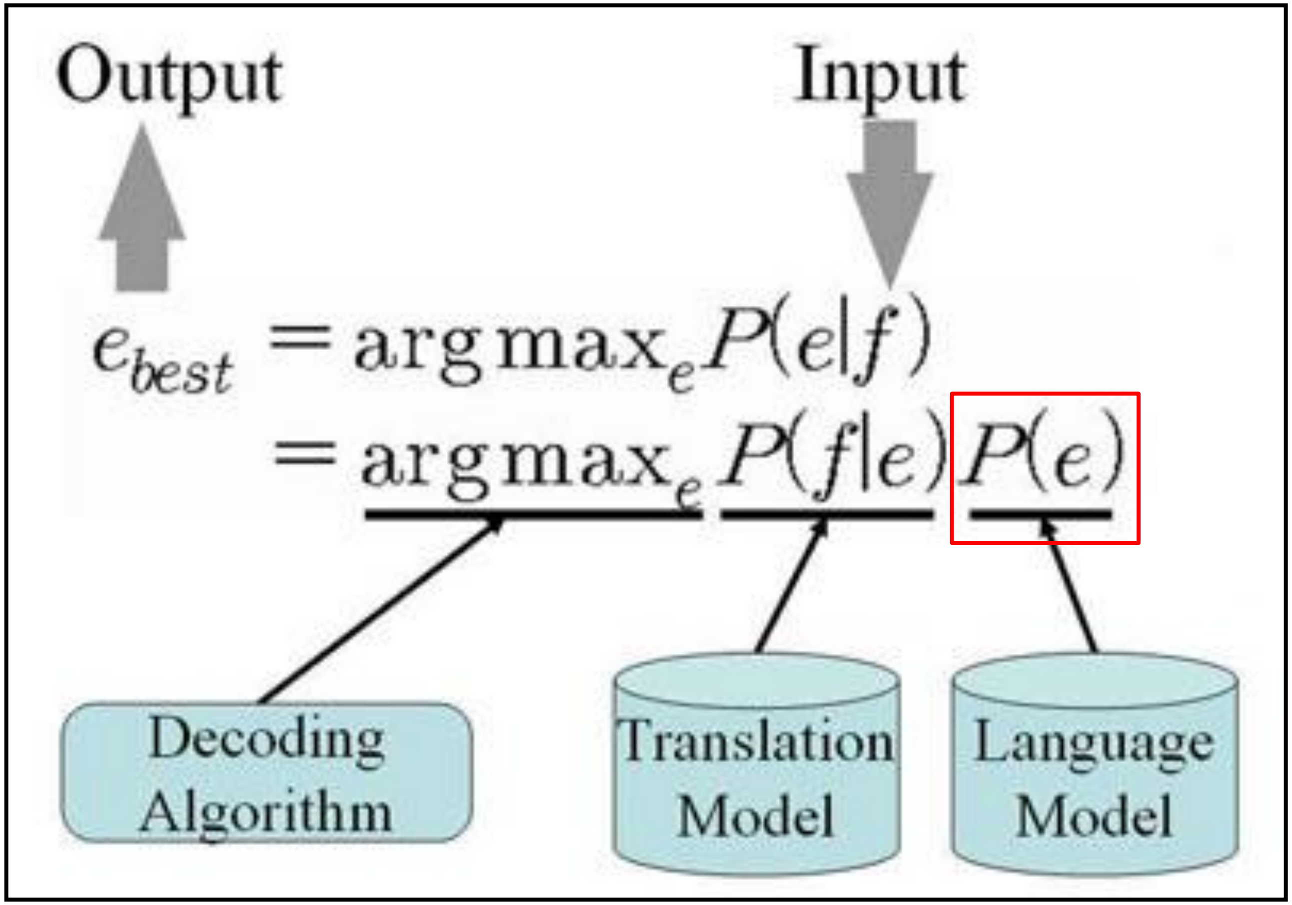
Based on Kevin Knight's 1999
A Statistical MT Tutorial Work Book
and some slides from Philipp Koehn

Overview

- Introduction: the basic idea
- IBM models: the noisy channel, Model 3, EM
- Language Models: the basic idea
- Phrase-Based SMT



Translation Modelling



Language Models

- Core component in SMT

- The IBM 3 (and other) SMT translation models $P(f|e)$ can be complex
- A lot can go wrong = the translation model can (and will!) produce lots of strange looking e's
- Of course we hope that the probabilities for the parameters used in modelling $P(f|e)$ will produce some good ones …
- Still we need a bit more help …

- The Language Model
- Recall: $\hat{e} = \arg max_e P(e|f) = \arg max_e P(f|e) \times P(e)$
- $P(e)$ trained on good English text (mono-lingual)

Language Models

$$\hat{e} = \arg max_e\, P(e|f) = \arg max_e\, P(f|e) \times P(e)$$

- Just an aside:
- If we'd reason directly about $P(e|f)$ (rather than go through noisy channel model with Baysian inversion) our probability estimates better be very good!

$$P(e|f) = \frac{P(e,f)}{P(f)}$$

- Going through noisy channel + inversion allows $P(f|e)$ to be a bit more lax/crazy (and easier to build) as it is being kept in check by $P(e)$.

Language Models

- What is $P(e)$?
- … the probability of English sentences

- E.g. suppose we have a million $(1{,}000{,}000)$ English sentences
- Suppose the sentence "*How's it going?* " occurs 56 times in the data
- Then we could use MLE to estimate

$$P(\text{How's it going?}) = \frac{56}{1{,}000{,}000}$$

- That seems reasonable …

Language Models

- Is this reasonable?
- Do we only want to look at full sentences in the data?
- And do we only assign $P(e)$ to grammatically correct sentences?

- No!

- We'll never see all possible English sentences in the data
- So a perfectly good but unseen sentence will just get $P(e) = 0$
- Also $P(f|e)$ will produce quite a bit of junk and even the best may not be 100% grammatical
- People sometimes say things that are ungrammatical …

Language Models

UNIVERSITÄT DES SAARLANDES

- So how do we build these models?
- We break things (sentences) down into (sequences) of words = n-grams
- Use these as building blocks for our $P(e)$ model

single word	unigram
two words	bigram
three words	trigram
…	…
n words	n-gram

- Idea: if a string has many reasonable n-grams, it is possibly ok.

Language Models

- A first go: let's use the chain rule

$$P(w_1 w_2 w_3 \ldots w_n) \\ = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1 w_2) \times \cdots \times P(w_n|w_1 \ldots w_{n-1})$$

- Not bad. Each of the parameters of the factors can be estimated using MLE with counts from large sets of mono-lingual data:

$$b(w) = \frac{\#(w)}{total\ word\ tokens\ in\ data} \qquad b(w_i|w_j) = \frac{\#(w_j w_i)}{\# w_j}$$

$$b(w_i|w_k \ldots w_j) = \frac{\#(w_k \ldots w_j w_i)}{\#(w_k \ldots w_j)}$$

- Trouble: many longer $b(w_i|w_k \ldots w_j)$ will never be seen in data

Language Models

- Problem is estimating these long "histories"
- Apply the Markov assumption: limited history/memory
- Instead of

$$P(w_1 w_2 w_3 \dots w_n) \\ = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1 w_2) \times \cdots \times P(w_n|w_1 \dots w_{n-1})$$

- We do sth. like

$$P(w_1 w_2 w_3 \dots w_n) \\ \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times \cdots \times P(w_n|w_{n-1})$$

- Decompose string into sequences of bigrams
- Often with "invisible" beginning and end of sentence marker:

$$P(w_1 w_2 w_3 \dots w_n) \\ \approx P(w_1|\langle s \rangle) \times P(w_2|w_1) \times P(w_3|w_2) \times \cdots \times P(w_n|w_{n-1}) \\ \times P(\langle /s \rangle|w_n)$$

Language Models

$$P(w_1 w_2 w_3 \dots w_n) \approx P(w_1|\langle s\rangle) \times P(w_2|w_1) \times P(w_3|w_2) \times \cdots \times P(w_n|w_{n-1}) \times P(\langle /s\rangle|w_n)$$

- Bigram LM, first-order Markov Model:

$$P(w_1 w_2 w_3 \dots w_n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1})$$

- Trigram LM, second-order Markov Model:

$$P(w_1 w_2 w_3 \dots w_n) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}\, w_{i-1})$$

Language Models

- Let $b(y|x)$ be the probability that word y follows word x

- b is a parameter of a generative probabilistic model that generates English strings and assigns probabilities to them

- We need to estimate b from data: lots of English text
- Using MLE, an estimator could look like this

$$b(y|x) = \frac{count("x\ y")}{count("x")} = \frac{\#("x\ y")}{\#("x")}$$

Language Models

$$P(I\ like\ snakes\ that\ are\ not\ poisonous\ .) = ?$$

$$\begin{aligned} P(I\ like\ snakes\ that\ are\ not\ poisonous\ .) \approx \\ & P(I|\text{start}-\text{of}-\text{sentence}) \times \\ & P(like|I) \times \\ & P(snakes|like) \times \\ & P(that|snakes) \times \\ & P(are|that) \times \\ & P(not|are) \times \\ & P(poisonous|not) \times \\ & P(.|poisonous) \times \\ & P(\text{end}-\text{of}-\text{sentence}|.) \end{aligned}$$

Language Models

$$P(How's\ it\ going\ ?) = ?$$

$$\begin{aligned} P(How\ 's\ it\ going\ ?) \approx\ & \\ & P(How|\text{start}-\text{of}-\text{sentence}) \times \\ & P('s|How) \times \\ & P(it|'s) \times \\ & P(going|it) \times \\ & P(?\,|going) \times \\ & P(\text{end}-\text{of}-\text{sentence}|?) \end{aligned}$$

- This is a bigram model
- Only remembers the previous word …

Language Models

- Trigram model:

$$b(z|x\ y) = \frac{\#(x\ y\ z)}{\#(y\ z)}$$

$$\begin{aligned} P(How\ 's\ it\ going\ ?) \approx & \\ & P(How|\text{start}-\text{of}-\text{sentence start}-\text{of}-\text{sentence}) \times \\ & P('s|\text{start}-\text{of}-\text{sentence}\ \ How) \times \\ & P(it|How\ \ 's) \times \\ & P(going|'s\ \ it) \times \\ & P(?\,|it\ going) \times \\ & P(\text{end}-\text{of}-\text{sentence}|going\ ?) \times \\ & P(\text{end}-\text{of}-\text{sentence}\ \ \text{end}-\text{of}-\text{sentence}|?) \end{aligned}$$

Language Models

- N-gram models can assign probabilities to sentences they have never seen
- By piecing things together from n-grams
- They generalise much better to unseen data than direct estimation of complete sentences from data
- But: they can also assign probability 0 to some perfectly good sentences:
- A bi-gram model will assign a sentence probability 0 if there is at least one single bi-gram in the sentence it never saw in training: if y never followed x in our training data, then $P(y|x) = 0 \ldots$
- Same for trigram models: if z never followed $x\ y$ in our training data, then $P(z|x\ y) = 0 \ldots$

Language Models: Smoothing

UNIVERSITÄT DES SAARLANDES

- Instead of

$$b(z|x\ y) = \frac{\#(x\ y\ z)}{\#(x\ y)}$$

- We can use sth. like

$$b(z|x\ y) =$$

$0.95 \times \frac{\#(x\ y\ z)}{\#(x\ y)} +$	(trigram)
$0.04 \times \frac{\#(y\ z)}{\#(y)} +$	(bigram)
$0.008 \times \frac{\#(z)}{\#(words)} +$	(unigram)
0.002	(if all else fails)

- Note: (i) this assigns non-zero probabilities to all strings (also ungrammatical strings); (ii) smoothing coefficients sum to 1
- This is not (!) the best way! There is a lot more to LMs than we can cover here …!

Language Models: Training

UNIVERSITÄT DES SAARLANDES

- How do we estimate b parameters?
- Just count n-grams in large data sets and divide …
- Fairly easy … but you need to be a bit careful to scale this to very large data sets
- Consistent and sensible tokenisation …

Language Models: Evaluation

- Model
- Generative story + parameter values

- How do we know one model is better than another?

- One way to compare them: select some new $testdata$; what is the probability of a $model$ given the $testdata$?

$$P(model|testdata)$$

- Apply Bayes

$$P(model|testdata) = \frac{P(testdata|model) \times P(model)}{P(testdata)}$$

Language Models: Evaluation

UNIVERSITÄT DES SAARLANDES

$$P(model|testdata) = \frac{P(testdata|model) \times P(model)}{P(testdata)}$$

- The best model is the one that maximises $P(model|testdata)$.
- $P(testdata)$ is the same here
- Assume $P(model)$ is the same too
- Then the best model is the one that maximises $P(testdata|model)$
- $P(testdata|model)$ is easy to compute:

$$P(testdata|model) = P(e)\text{, where } e = testdata$$

Language Models: Evaluation

UNIVERSITÄT DES SAARLANDES

- Trigram models generally better than bigram
- A test sentence like

I hire men who is good pilots

- Will get fairly high probability by bigram model

$$b(who|men)\ \ b(is|who)$$

- But not by trigram model

$$b(is|men\ who)$$

Language Models: Evaluation

UNIVERSITÄT DES SAARLANDES

- Perplexity per word: N is number of words in $e = w_1 w_2 \dots w_N$

$$PPL = 2^{-\frac{\log_2 P(e)}{N}} = 2^{-\frac{1}{N}\log_2 P(e)}$$

- If a model assigns high $P(e)$ to some unseen data e, it is not very surprised by the data and perplexity is low
- As $P(e)$ increases, perplexity decreases
- Better models have lower perplexity
- $-\log_2 P(e)$ optimal (= minimal) number of bits to code e

Language Models: Evaluation Example

$$2^{-\frac{\log_2 P(e)}{N}} = 2^{-\frac{1}{N}\log_2 P(e)}$$

- Suppose we have a unigram language model with

$$p(x) = \frac{1}{4}, p(y) = \frac{1}{2}, p(z) = \frac{1}{4}$$

- What is the perplexity of the string "$x\ y\ z$" ?

$P(e) = \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{32}$ and $\log_2 P(e) = \log_2(\frac{1}{32}) = -5$

$-\frac{1}{N}\log_2 P(e) = -\frac{1}{3} - 5 = \frac{5}{3}$ and $2^{-\frac{1}{N}\log_2 P(e)} = 2^{\frac{5}{3}} \approx 3.175$

Language Models

$$2^{-\frac{\log_2 P(e)}{N}} = 2^{-\frac{1}{N}\log_2 P(e)}$$

$$\log_2 P(e)$$

$$P(e) = P(w_1 w_2 \dots w_n)$$
$$= P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_n|w_1 \dots w_{n-1})$$

$$\log_2 P(e) = \log_2 P(w_1 w_2 \dots w_n)$$
$$= \log_2 (P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_n|w_1 \dots w_{n-1}))$$
$$= \log_2 P(w_1) + \log_2 P(w_2|w_1) + \cdots + \log_2 P(w_n|w_1 \dots w_{n-1})$$
$$= \sum_{i=1}^{n} \log_2 P(w_i|w_1 \dots w_{i-1})$$

Language Models

$$2^{-\frac{\log_2 P(e)}{N}} = 2^{-\frac{1}{N}\log_2 P(e)} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_1 \ldots w_{i-1})}$$

$$\left(P(w_1 w_2 \ldots w_n)\right)^{-\frac{1}{n}} = \sqrt[n]{\frac{1}{P(w_1 w_2 \ldots w_n)}}$$

$$a^x = e^{\ln(a^x)} = e^{x\ln(a)}$$

- Perplexity Intuition: average number of choices/branching factor

Evaluating (S)MT

- What do we want to know?
  - How good is the (S)MT output?
  - Is a system useful?
  - Is one system better than another?

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

- How many good translations are there?

Evaluating (S)MT

- Why do we want to evaluate (S)MT?

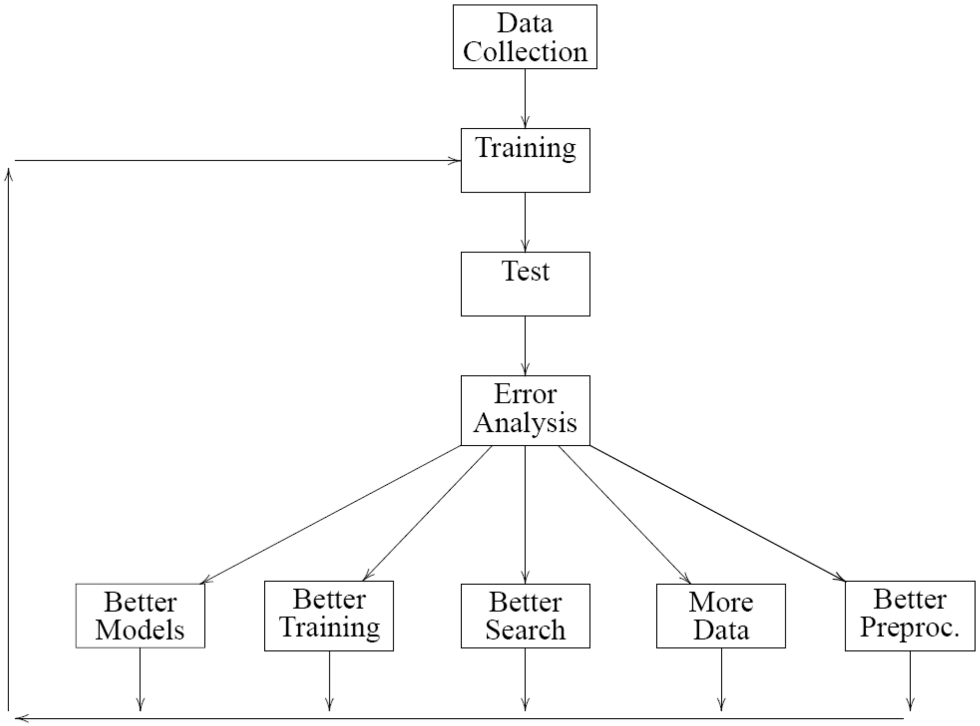


Figure 3.1: Development cycle of a statistical MT system.

(Och 2000)

Evaluating (S)MT

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

- Manual
  - Human professional translators
  - People proficient in source and target language at stake
  - People who only understand target but have access to a reference?
  - Can be time consuming and expensive
  - Not easy to reproduce: rater/inter-annotator agreement
  - MT output sometimes so bad, hard to rate …
  - Still: the yardstick, the gold-standard, the ideal …

Human Evaluation

- Guidelines
- Adequacy (Scale of 5)
  1. All meaning
  2. Most meaning
  3. Much meaning
  4. Little meaning
  5. none
- Fluency (Scale of 5)
  1. Flawless (English)
  2. Good (English)
  3. Non-native (English)
  4. Disfluent (English)
  5. Incomprehensible

Human Evaluation

UNIVERSITÄT DES SAARLANDES

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 .

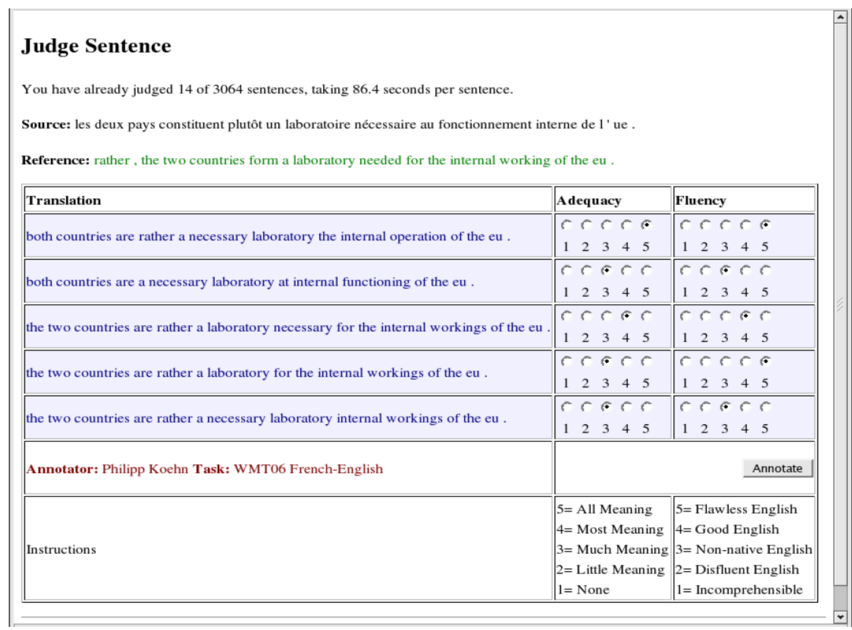
Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

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

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 …
- Miserable inter-annotator/rater agreement
- Don't know what is wrong or 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 .

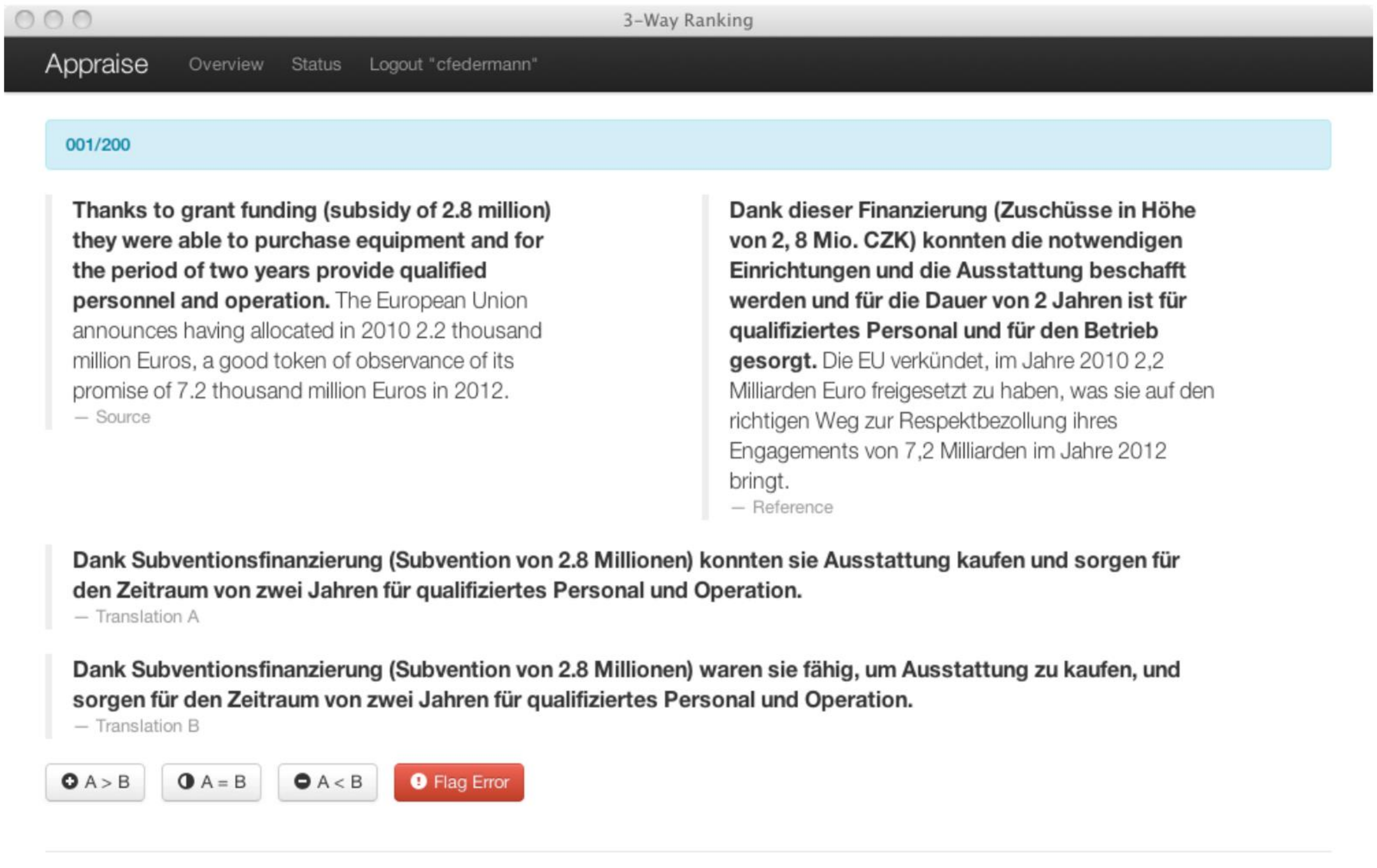
Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

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

UNIVERSITÄT DES SAARLANDES

Appraise (Christian Federmann 2012)



Human Evaluation

UNIVERSITÄT DES SAARLANDES

- Error classifications (Vilar et al. 2006):

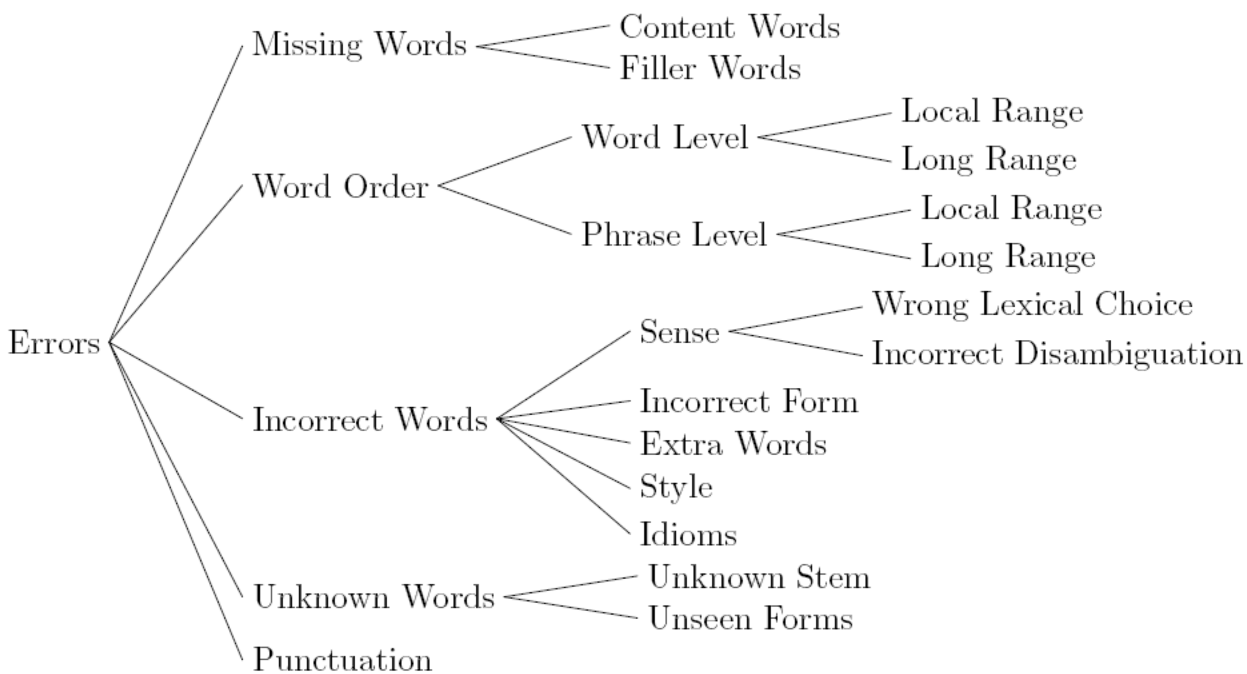
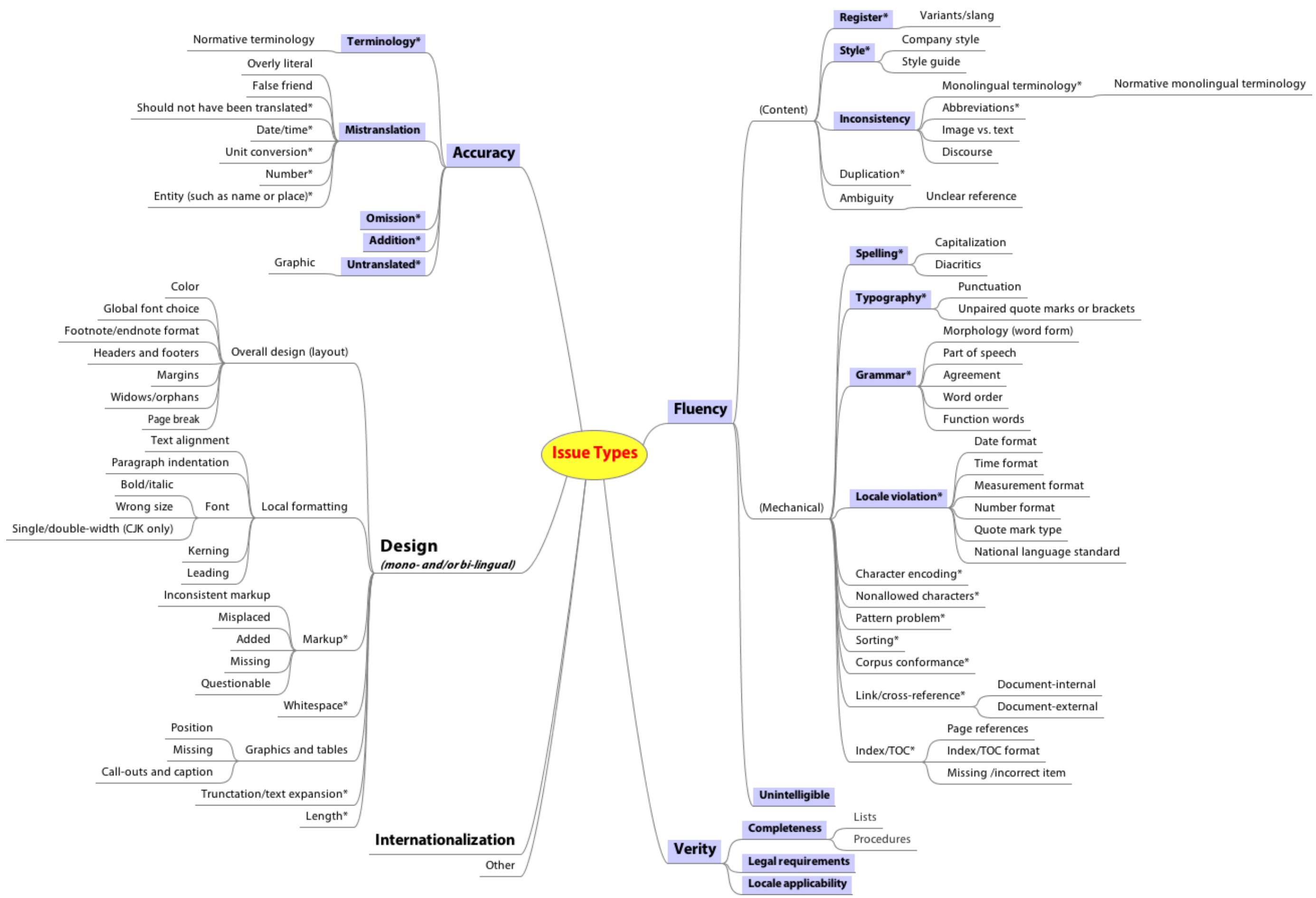


Figure 1: Classification of translation errors.

Human Evaluation

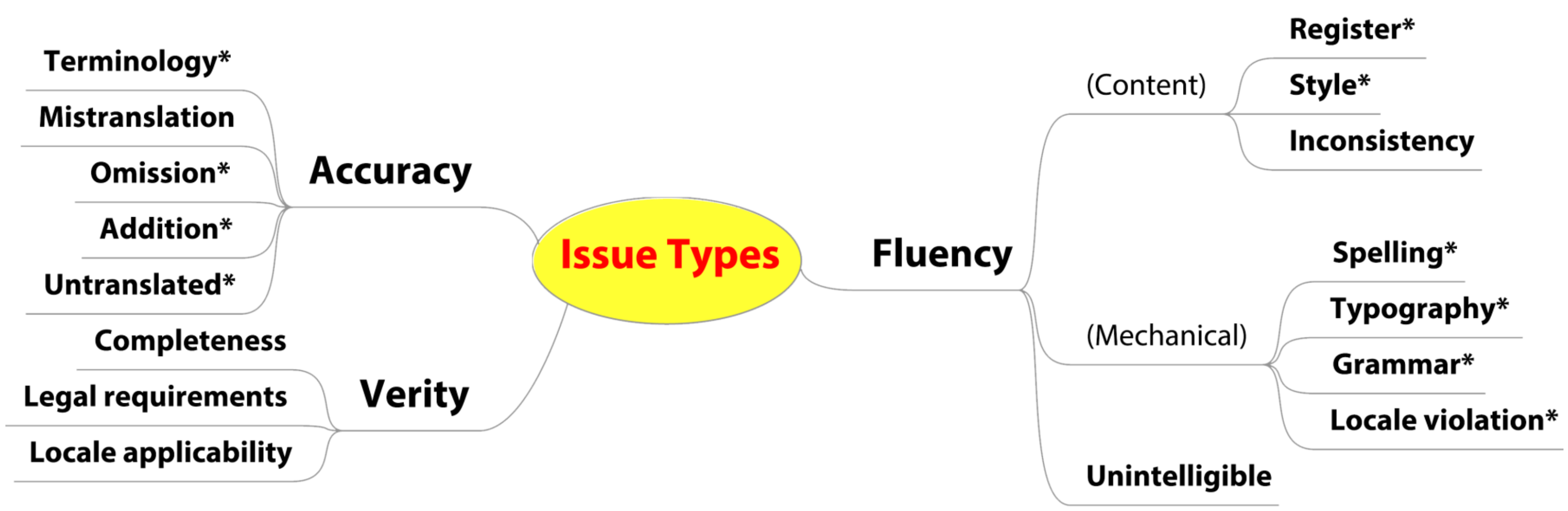
UNIVERSITÄT DES SAARLANDES

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



Human Evaluation

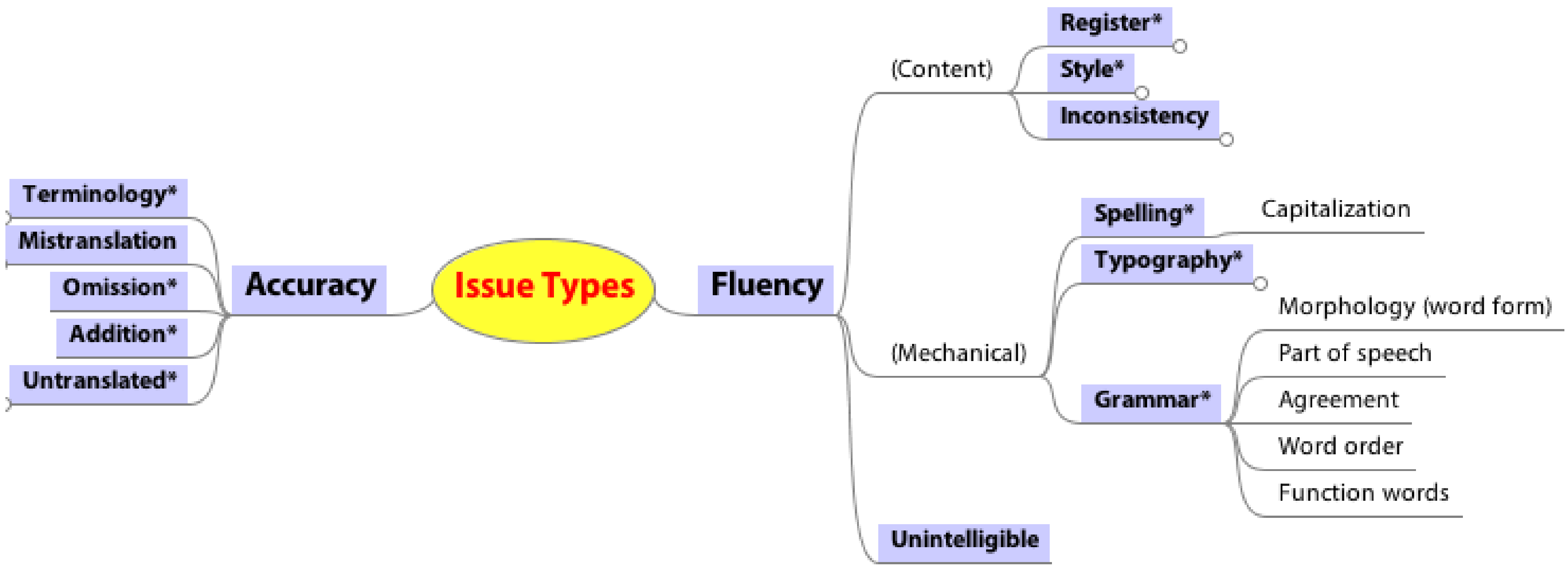
- Error classification MQM Core QTLaunchPad (Lommel et al. 2013):



Human Evaluation

UNIVERSITÄT DES SAARLANDES

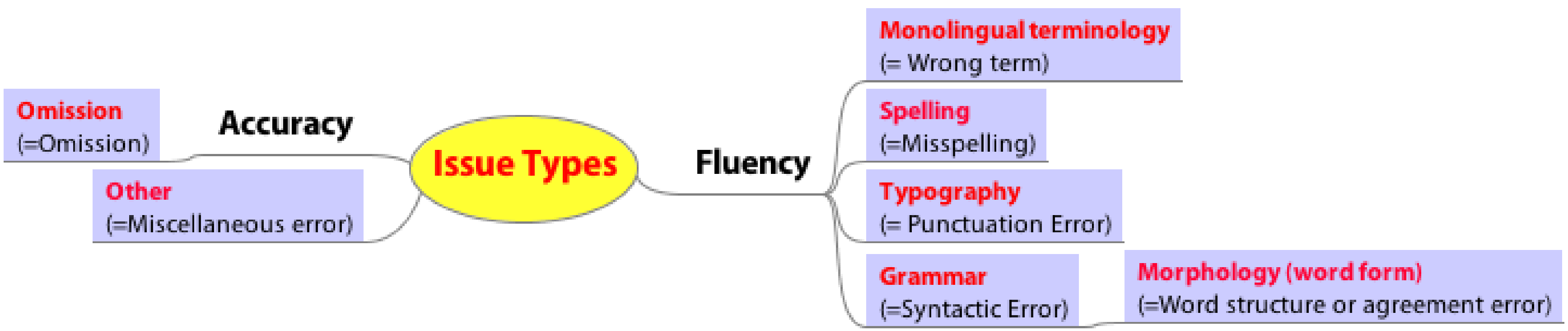
- Error classification MQM MT Subset (Lommel et al. 2013):



Human Evaluation

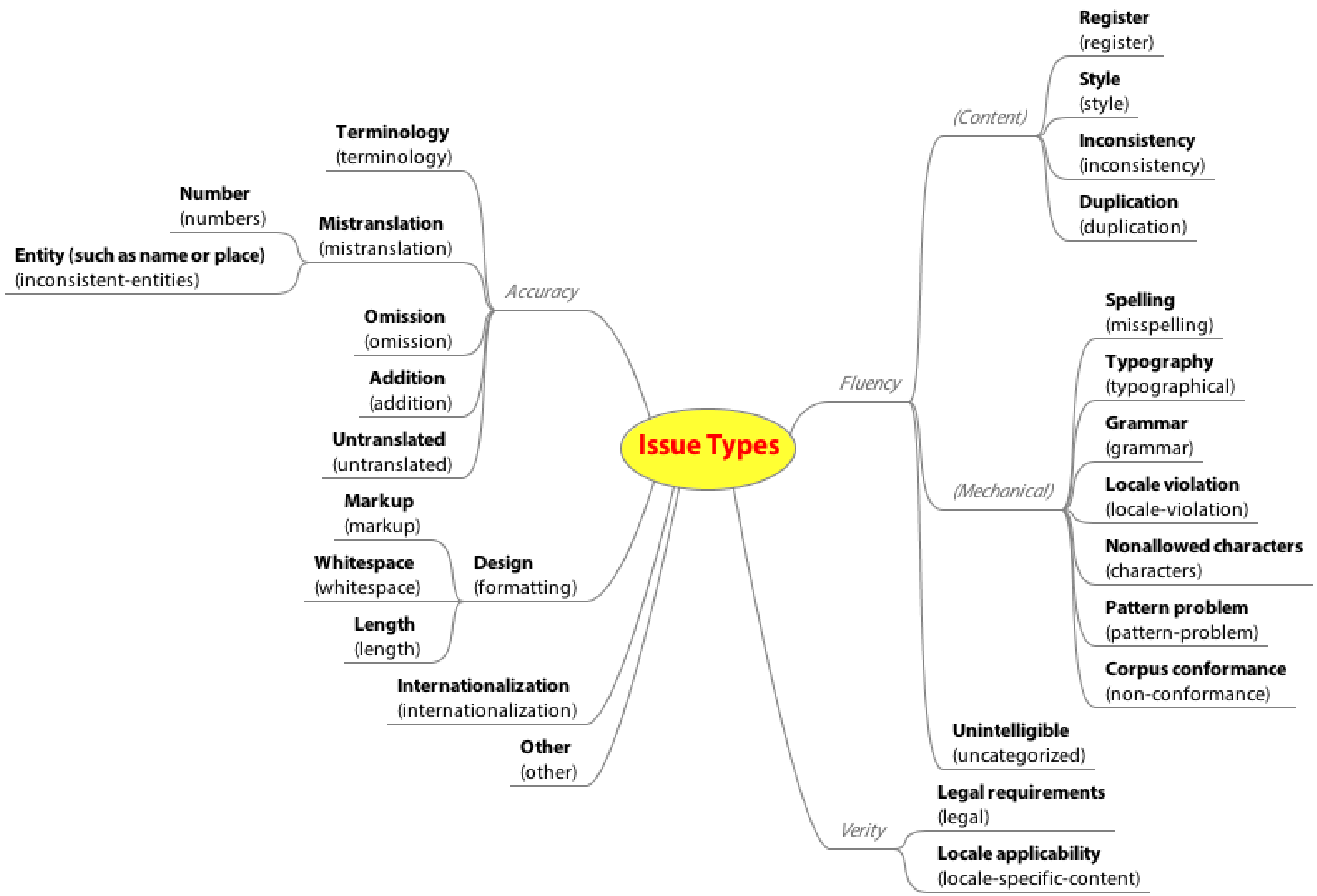
UNIVERSITÄT DES SAARLANDES

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



Human Evaluation

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



Human Evaluation:

- 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)
- 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 …

Automatic Evaluation

UNIVERSITÄT DES SAARLANDES

- The basic idea
- Given a reference translation (or several reference translations), compare MT output against
- How?
- How similar are they?
- Word, n-gram, string-overlap (surface string similarity)
- More sophisticated stuff (not just surface string matching based)
  - Stemming, morphological analysis, synonyms, paraphrases, syntactic and semantic structure, etc.

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 of the words in output are correct?

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

- 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

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	0.86	0.86

- Problem: f-measure can reward unintelligible word salad 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 words
- n-gram precision $(n = 1 \ldots 4)$, plus brevity penalty

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

- $BLEU = 0$ if the hypothesis does not have a matching n-gram for any of the $n = 1 \ldots 4$: System 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	0.86	0.86
BLEU	0	0.52	0

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

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:

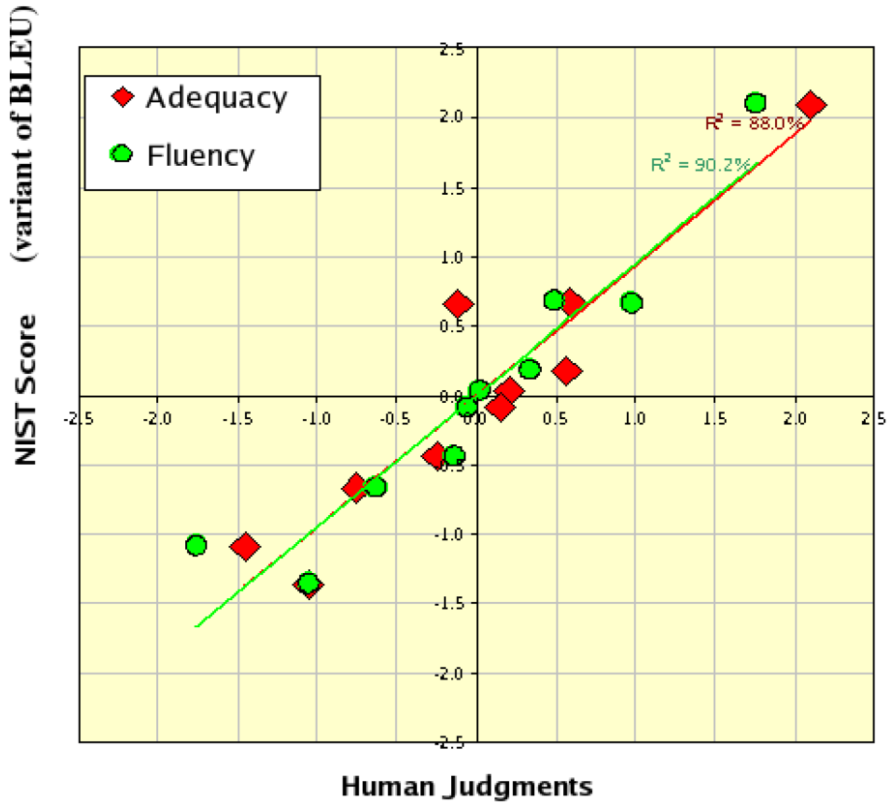
$$BLEU = \min(1, \exp(1 - \frac{|reference|}{|output|}))(\exp\left(\sum_{n=1}^{4} \lambda_n \times \log(n - gram\ prec)\right))^{\frac{1}{4}}$$

- λ_i usually 1…

Automatic Evaluation: BLEU

UNIVERSITÄT DES SAARLANDES

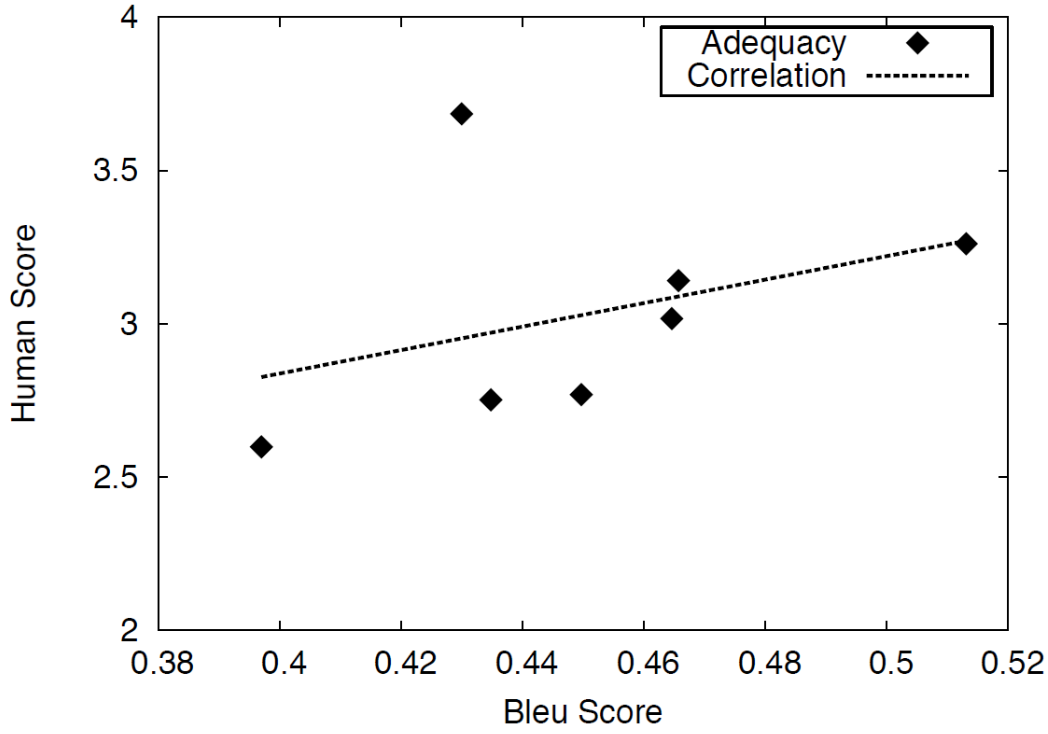
Correlation with Human Judgement



Automatic Evaluation: BLEU

Evidence of Shortcomings of Automatic Metrics

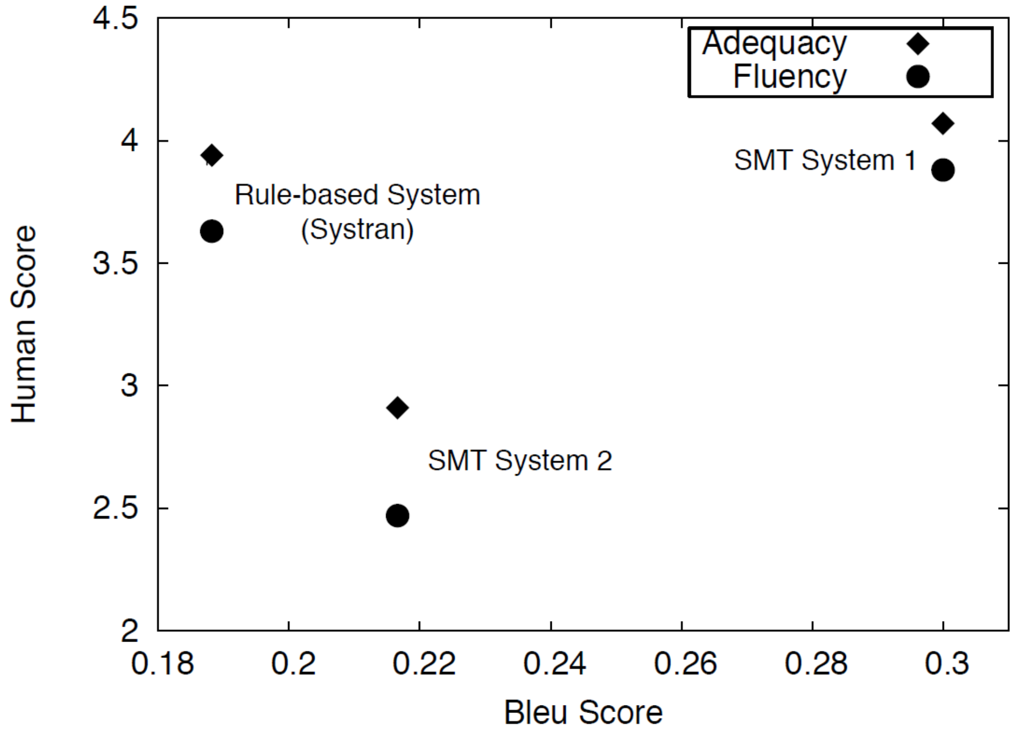
Post-edited output vs. statistical systems (NIST 2005)



Automatic Evaluation: BLEU

Evidence of Shortcomings of Automatic Metrics

Rule-based vs. statistical systems



Automatic MT Evaluation Metrics

UNIVERSITÄT DES SAARLANDES

- Treat all words as strings: no difference between function and content words
- Do not consider global grammaticality
- Do not consider meaning

Yesterday John resigned from the company

John quit the company yesterday

- Scores by themselves do not mean much
- Human translators score low on BLEU

- But: many references
- METEOR, MEANT, Karolina Owczarzack …