Metrics of MT Quality

Ondřej Bojar

February 20, 2020




Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Course Outline

1. **Metrics of MT Quality.**
2. Approaches to MT. SMT, PBMT, NMT, NP-hardness.
3. NMT (Seq2seq, Attention. Transformer). Neural Monkey.
4. Parallel texts. Sentence and word alignment. hunalign, GIZA++.
5. PBMT: Phrase Extraction, Decoding, MERT. Moses.
6. Morphology in MT. Factors or segmenting, data or linguistics.
7. Syntax in SMT (constituency, dependency, deep).
8. Syntax in NMT (soft constraints/multitask, network structure).
9. Towards Understanding: Word and Sentence Representations.
10. Advanced: Multi-Lingual MT. Multi-Task Training. Chef's Tricks.
11. Project presentations.

Outline

- Task of MT (formulating a simplified goal).
- Manual evaluation.
- Automatic evaluation.
- Empirical confidence bounds.
- End-to-end vs. component evaluation.
- Summary: Evaluation caveats.

Importance of Measuring MT Output

You need a metric to be able to check your progress.
An example from the history:

- Manual judgement at Euratom (Ispra) of a Systran system (Russian→English) in 1972 revealed huge differences in judging; (Blanchon et al., 2004):
  - 1/5 (D–) for output quality (evaluated by teachers of language),
  - 4.5/5 (A+) for usability (evaluated by nuclear physicists).

Metrics can drive the research for the topics they evaluate.

- Some measured improvement required by sponsors: NIST MT Eval, DARPA, TC-STAR, EuroMatrix+.
- BLEU has lead to a focus on phrase-based MT.
- Other metrics may similarly change the community's focus.

Our Goal in MT

We restrict the task of MT to the following conditions.

- No writers' ambitions, we prefer literal translation.
- No attempt at handling cultural differences.

Expected output quality:

1. Worth reading. (Not speaking the src. lang. I can sort of understand.)
2. Worth editing. (I can edit the MT output to obtain publishable text.)
3. Worth publishing, no editing needed.

In general, we're aiming at level 1 or 2. Level 3 remains risky.

Basic Manual Evaluation Decisions

What to Show to the annotators when assessing the candidate?

- REF-based ... only the (human) reference
- SRC-based ... only the source
- SRC&REF-based ... both

Context to Consider:

- Sentence-level ... sentences in random order
- Document-level ... obtain single score per document
- Document-aware ... show whole documents, scores per sentence

What to Ask from annotators (scoring technique):

- Some relative score over several candidates?
- Some absolute score for a single output?
- A more complicated question?

Scoring Techniques

Black-box: Judging hypotheses produced by MT systems:

- Adequacy and fluency of whole sentences.
  Somewhat revisited under the name Direct assessment (DA).
- Relative ranking (RR) of full sentences by several MT systems:
  Longer sentences hard to rank. Candidates incomparably poor.
- Ranking of constituents, i.e. parts of sentences:
  Tackles the issue of long sentences. Does not evaluate overall coherence.
- Comprehension test: Blind editing+correctness check.
- Task-based: Does MT output help as much as the original?
  Do I dress appropriately given a translated weather forecast?

Gray-box: Analyzing errors in systems' output.

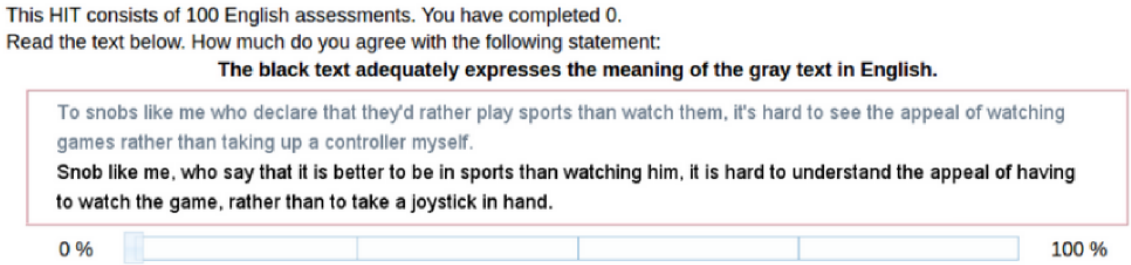
- HMEANT, HUME: Is the core event structure preserved?
- MQM: Multi-dimensional quality metrics.

Glass-box: System-dependent: Does this component work?

Direct Assessment: Adequacy

Graham et al. (2013) propose a simple continuous scale:

- To what extent MT adequately expresses the meaning of REF?

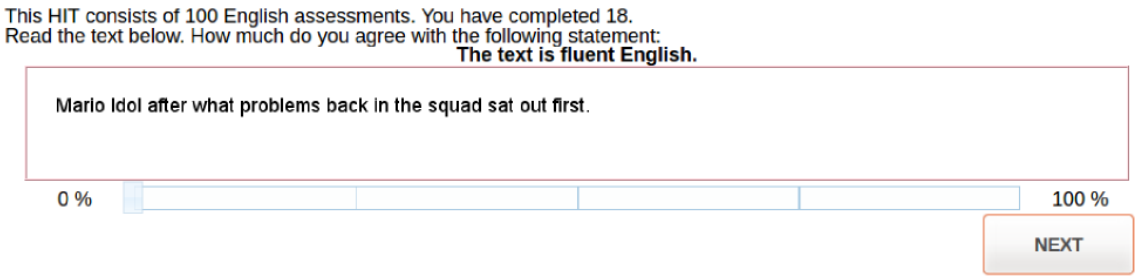


⊕ After ∼15 judgements, each annotator stabilizes.
⊖ Interpretable by averaging over many judgements of many people.
⊖ 30–70(!)% of participating Turkers unreliable.
⊖ Too few non-English speakers on Amazon Mechanical Turk.

Direct Assessment: Fluency

DA for fluency:

- To what extent the MT is fluent English?
- The source or reference are not shown at all.
- Fluency used only to break ties in adequacy.



Recent Result: MT Surpassing Humans: 2018

- WMT 2018 English-to-Czech news translation results: (Bojar et al., 2018)

	Ave. %	Ave. z	System
1	84.4	0.667	**CUNI-Transformer**
2	79.8	0.521	UEDIN
	78.6	0.483	Professional Translation
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

Recent Result: MT Surpassing Humans: 2018

- WMT 2018 English-to-Czech news translation results: (Bojar et al., 2018)

	Ave. %	Ave. z	System
1	84.4	0.667	**CUNI-Transformer**
2	79.8	0.521	UEDIN
	78.6	0.483	Professional Translation
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

Caveats:

- Humans translated *whole documents*, MT individual segments.
  - Evaluation was done for *individual segments*.

SRC-Based Doc-Level DA

- ⊖ Mental overload.
- ⊖ Too few scores collected ⇒ Difficult to get statistical significance.

SRC-Based Pseudo-Doc-Aware DA

- Score sentences using DA one by one.
- *In the original order* (i.e. not shuffled).

$\Rightarrow$ Mentally manageable.

Problems of the first run at WMT19 (Barrault et al., 2019):

- No way to go back to previous sentences.
- All sentences in a row must come from the same MT system.
- No longer independent probes (violating statistical assumptions).

Recent Results: MT Surpassing Humans: 2019

English→Czech

	Ave.	Ave. z	System
1	91.2	0.642	Professional Translators
2	86.0	0.402	CUNI-DocTransformer-T2T
	86.9	0.401	CUNI-Transformer-T2T-2018
	85.4	0.388	CUNI-Transformer-T2T-2019
5	81.3	0.223	CUNI-DocTransformer-Marian
	80.5	0.206	uedin
7	70.8	−0.156	online-Y
	71.4	−0.195	TartuNLP-c
9	67.8	−0.300	online-G
10	68.0	−0.336	online-B
11	60.9	−0.594	online-A
12	59.3	−0.651	online-X

Recent Results: MT Surpassing Humans: 2019

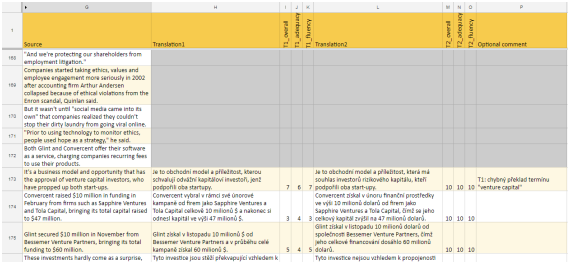
English→Czech

	Ave.	Ave. z	System
1	91.2	0.642	Professional Translators
2	86.0	0.402	CUNI-DocTransformer-T2T
	86.9	0.401	CUNI-Transformer-T2T-2018
	85.4	0.388	CUNI-Transformer-T2T-2019
5	81.3	0.223	CUNI-DocTransformer-Marian
	80.5	0.206	uedin
7	70.8	−0.156	online-Y
	71.4	−0.195	TartuNLP-c
9	67.8	−0.300	online-G
10	68.0	−0.336	online-B
11	60.9	−0.594	online-A
12	59.3	−0.651	online-X

English→German

Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	Professional Translation
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
		... 10 more systems here ...
76.3	−0.400	online-X
43.3	−1.769	en-de-task

SRC-Based Doc-Aware 10-RankME



	Source	Translation1	T1_overall	T1_adequacy	T1_fluency	Translation2	T2_overall	T2_adequacy	T2_fluency	Optional comment
168	"And we're protecting our shareholders from employment litigation."									
169	Companies started taking ethics, values and employee engagement more seriously in 2002 after accounting firm Arthur Andersen collapsed because of ethical violations from the Enron scandal, Quinlan said.									
170	But it wasn't until "social media came into its own" that companies realized they couldn't stop their dirty laundry from going viral online.									
171	"Prior to using technology to monitor ethics, people used hope as a strategy," he said.									
172	Both Glint and Convercent offer their software as a service, charging companies recurring fees to use their products.									
173	It's a business model and opportunity that has the approval of venture capital investors, who have propped up both start-ups.	Je to obchodní model a příležitost, kterou schvalují odvážní kapitáloví investoři, jenž podpořili oba start-upy.	7	6	7	Je to obchodní model a příležitost, která má souhlas investorů rizikového kapitálu, kteří podpořili oba start-upy.	10	10	10	T1: chybný překlad termínu "venture capital"
174	Convercent raised $10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to $47 million.	Convercent vybral v rámci své únorové kampaně od firem jako Sapphire Ventures a Tola Capital celkově 10 milionů $ a navíc se odnesl kapitál ve výši 47 milionů $.	3	4	3	Convercent získal v únoru finanční prostředky ve výši 10 milionů dolarů od firem jako Sapphire Ventures a Tola Capital, čímž se jeho celkový kapitál zvýšil na 47 milionů dolarů.	10	10	10	
175	Glint secured $10 million in November from Bessemer Venture Partners, bringing its total funding to $60 million.	Glint získal v listopadu 10 milionů $ od Bessemer Venture Partners a v průběhu celé kampaně získal 60 milionů $.	5	4	5	Glint získal v listopadu 10 milionů dolarů od společnosti Bessemer Venture Partners, čímž jeho celkové financování dosáhlo 60 milionů dolarů.	10	10	10	
176	These investments hardly come as a surprise, given the interconnected nature of companies, culture and venture capital.	Tyto investice jsou stěží překvapující vzhledem k vzájemné povaze společností, kultury a rizikového kapitálu.	3	4	3	Tyto investice nejsou vzhledem k propojenosti společností, kultury a rizikového kapitálu žádným překvapením.	10	10	10	
177	There's a growing body of research showing today's employees expect more from their workplaces than before.	Narůstající počet výzkumů jasně potvrzuje, že dnešní zaměstnanci očekávají od svého pracoviště více než dříve.	5	5	8	Roste množství výzkumů, které ukazují, že dnešní zaměstnanci očekávají od svých pracovišť více než dříve.	10	10	10	
178	In competitive markets such as Silicon Valley, high salaries and interesting projects are merely table stakes.	A na konkurenčních trzích, jakým je např. Silicon Valley, nejsou hlavní výhodou vysoké platy a zajímavé projekty.	6	5	8	Na konkurenčních trzích, jako je Silicon Valley, jsou vysoké platy a zajímavé projekty pouhými sázkami u stolu.	7	8	7	problém: význam terminu "table stakes"
179	Employees want to feel that they're accepted and valued and that they're giving their time to a company with a positive mission.	Zaměstnanci chtějí vnímat, že jsou přijímáni a oceňováni a že věnují svůj čas společnosti, která usiluje o pozitivní poslání.	9	9	9	Zaměstnanci chtějí mít pocit, že jsou přijímáni a oceněni a že věnují svůj čas společnosti s pozitivním posláním.	10	10	9	

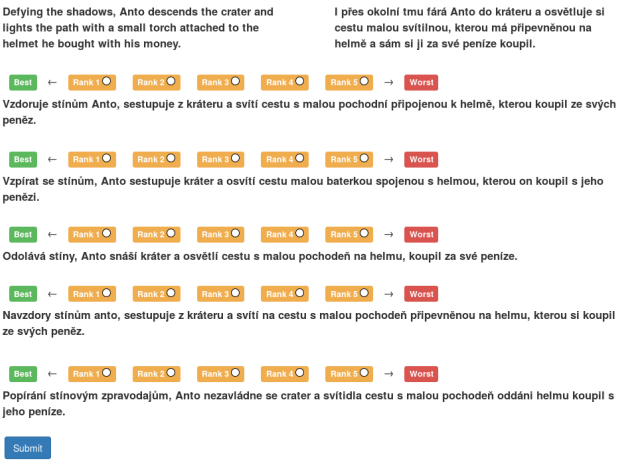
Mix of all:

- Two or more systems considered.
- Whole document shown.
- A section of 10 consecutive sentences scored in (1) adequacy, (2) fluency, (3) overall.

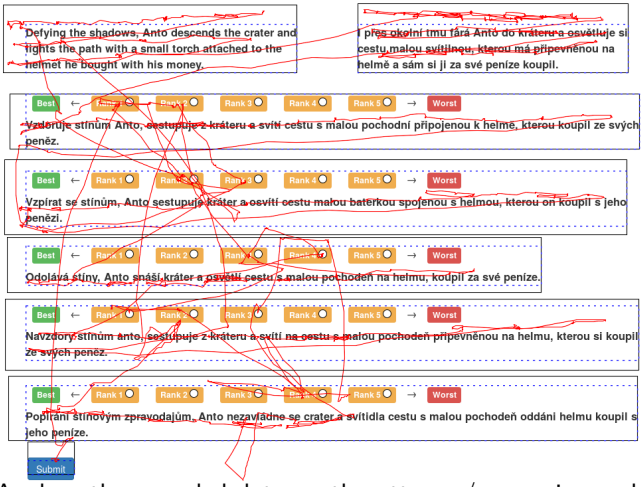
⇒ Combines relative, absolute, doc-level, sent-level.

⊖ Very time-consuming.

Relative Ranking of Sentences



Relative Ranking (Eye-Tracked)



Project suggestion: Analyze the recorded data: path patterns / errors in words.

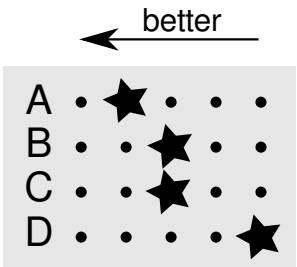
Relative Ranking of Constituents

Source: Können die USA **ihre Besetzung aufrechterhalten**, wenn sie dem irakischen Volk nicht Nahrung, Gesundheitsfürsorge und andere grundlegende Dienstleistungen anbieten können?

Reference: Can the US **sustain its occupation** if it cannot provide food, health care, and other basic services to Iraq's people?

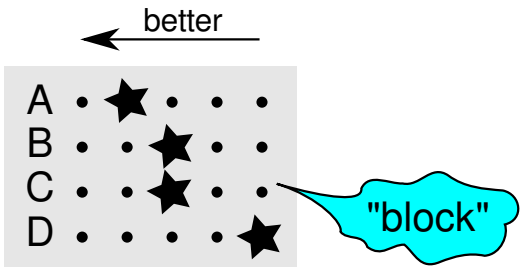
Translation	Rank
The United States can **maintain its employment** when it the Iraqi people not food, health care and other basic services on offer?.	○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 Worst Best
The US can **maintain its occupation**, if they cannot offer the Iraqi people food, health care and other basic services?	○ 1 ○ 2 ○ 3 ○ 4 ◉ 5 Worst Best
Can the US **their occupation sustained** if it to the Iraqi people not food, health care and other basic services can offer?	○ 1 ○ 2 ◉ 3 ○ 4 ○ 5 Worst Best
Can the United States **maintain their occupation**, if the Iraqi people do not food, health care and other basic services can offer?	○ 1 ○ 2 ○ 3 ◉ 4 ○ 5 Worst Best
The United States **is maintained**, if the Iraqi people, not food, health care and other basic services can offer?	○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 Worst Best
Annotator: ccb **Task:** WMT07 German-English News Corpus	
Instructions: Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the highlighted part** of each translation. *Please note that segments are selected automatically, and they should be taken as an approximate guide. They might include extra words on either end that are not in the actual alignment, or miss words.*	

Interpreting Manual Ranks



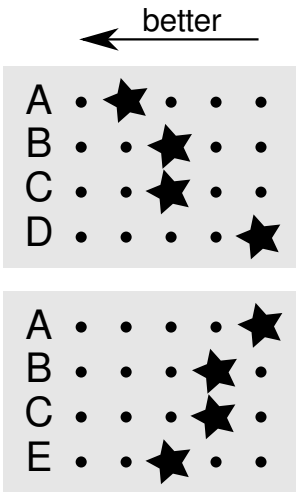
See also Bojar et al. (2011).

Interpreting Manual Ranks



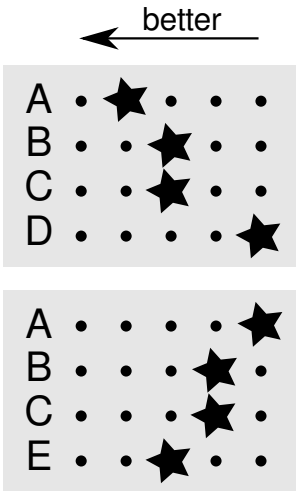
See also Bojar et al. (2011).

Interpreting Manual Ranks



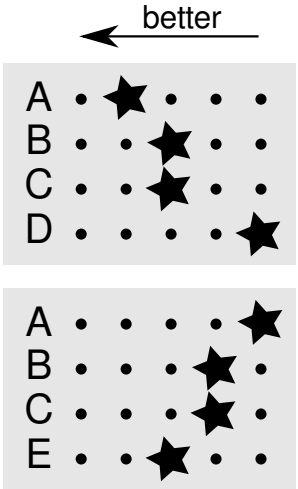
See also Bojar et al. (2011).

Interpreting Manual Ranks



See also Bojar et al. (2011).

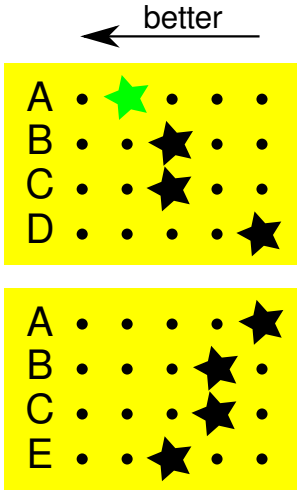
Interpreting Manual Ranks



[Systems] are ranked based on how frequently they were judged to be better than or equal to any other system.

See also Bojar et al. (2011).

Interpreting Manual Ranks

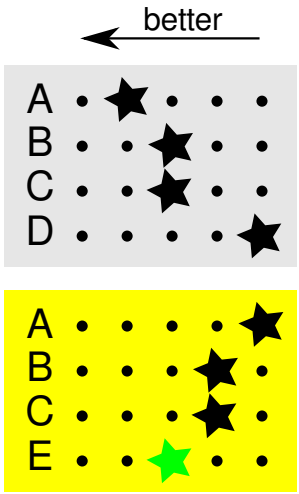


"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks

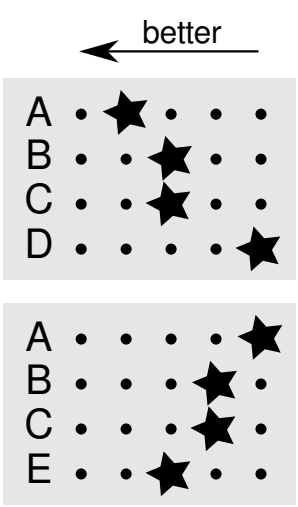


"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



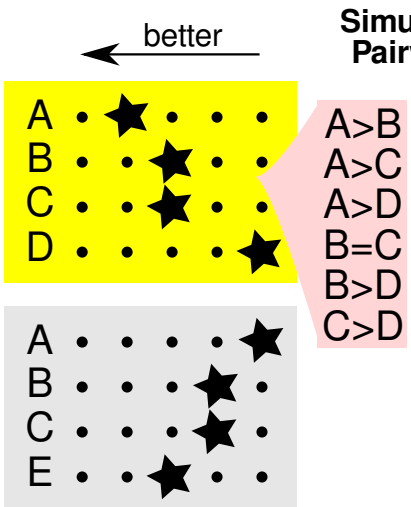
Simulated Pairwise

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

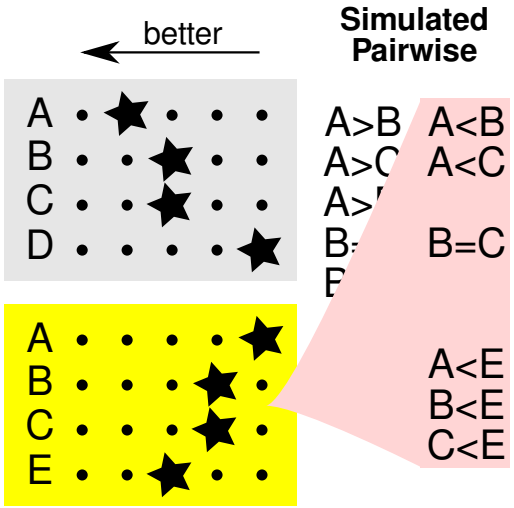
A>B
A>C
A>D
B=C
B>D
C>D

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks

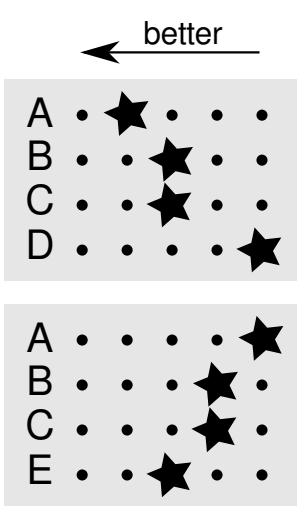


"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

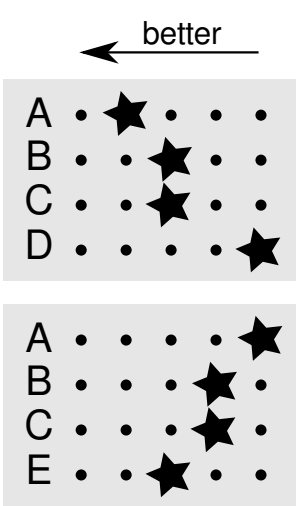
A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ Others"

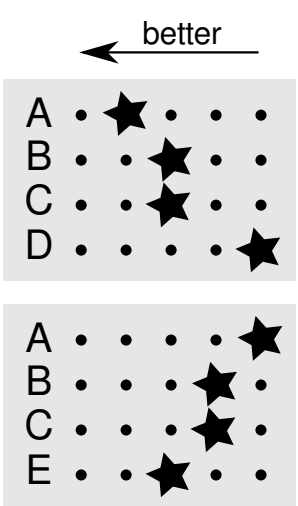
A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ Others"

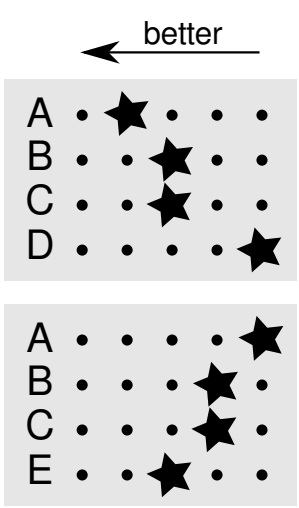
A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ Others"

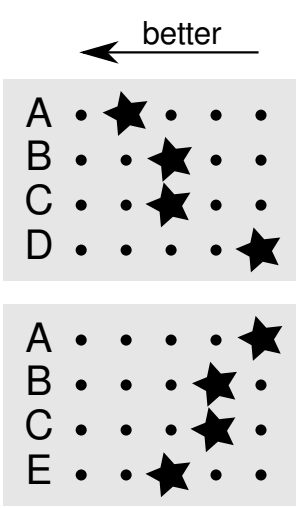
A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ Others"

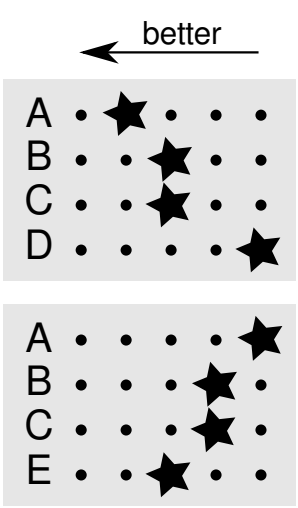
A: 3/6
B: 4/6
C: 4/6
D: 0/3
E: 3/3

"≥ All in Block"

A: 1/2
B: 0/2
C: 0/2
D: 0/1
E: 1/1

See also Bojar et al. (2011).

Interpreting Manual Ranks



Simulated Pairwise

A>B	A<B
A>C	A<C
A>D	
B=C	B=C
B>D	
C>D	
	A<E
	B<E
	C<E

"≥ Others"

A: 3/6

B: 4/6

"≥ All in Block"

A: 1/2

B: 0/2

See also Bojar et al. (2011).

Comprehension 1/2 (Blind Editing)

Original: They are often linked to other alterations sleep as nightmares, night terrors, the nocturnal enuresis (pee in bed) or the sleepwalking, but it is not always the case.

Edit:

They are often linked to other sleep disorders, such as nightmares, night terrors, the nocturnal enuresis (bedwetting) or sleepwalking, but this is not always the case.

Reset Edit

- (•) Edited.
- ( ) No corrections needed.
- ( ) Unable to correct.

Annotator: ccb **Task:** WMT09 Multisource-English News Editing

Instructions:
Correct the translation displayed, making it as fluent as possble. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."

Comprehension 2/2 (Judging)

Source: Au même moment, les gouvernements belges, hollandais et luxembourgeois ont en parti nationalisé le conglomérat européen financier, Fortis.
Les analystes de Barclays Capital ont déclaré que les négociations frénétiques de ce week end, conclues avec l'accord de sauvetage" semblent ne pas avoir réussi à faire revivre le marché".
Alors que la situation économique se détériorasse, la demande en matières premières, pétrole inclus, devrait se ralentir.
"la prospective d'équité globale, de taux d'intérêt et d'échange des marchés, est devenue incertaine" ont écrit les analystes de Deutsche Bank dans une lettre à leurs investisseurs."
"nous pensons que les matières premières ne pourront échapper à cette contagion.

Reference: Meanwhile, the Belgian, Dutch and Luxembourg governments partially nationalized the European financial conglomerate Fortis.
Analysts at Barclays Capital said the frantic weekend negotiations that led to the bailout agreement "appear to have failed to revive market sentiment."
As the economic situation deteriorates, the demand for commodities, including oil, is expected to slow down.
"The outlook for global equity, interest rate and exchange rate markets has become increasingly uncertain," analysts at Deutsche Bank wrote in a note to investors.
"We believe commodities will be unable to escape the contagion.

Translation	Verdict
While the economic situation is deteriorating, demand for commodities, including oil, should decrease.	◉ Yes ○ No
While the economic situation is deteriorating, the demand for raw materials, including oil, should slow down.	◉ Yes ○ No
Alors que the economic situation deteriorated, the request in rawmaterial enclosed, oil, would have to slow down.	○ Yes ◉ No
While the financial situation damaged itself, the first matters affected, oil included, should slow down themselves.	○ Yes ◉ No
While the economic situation is depressed, demand for raw materials, including oil, will be slow.	○ Yes ◉ No
Annotator: ccb **Task:** WMT09 French-English News Edit Acceptance	
Instructions: Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is **bold**.	

Quiz-Based Evaluation (1/1)

An approximation of task-based evaluation.

Preparation: English texts and Czech yes/no questions:

- We *found* English text snippets hopefully by native speakers.
- We equipped each snippet with 3 yes/no questions in Czech.

3 different snippet lengths (1..3 sents.), 4 different topics:

- Meeting: when, where, how often, with whom, ...
- Directions: driving/walking instructions, finding buildings, ...
- Basic quizes: maths, physics, biology, ... simple questions.
- Politics/News: elections chances, affairs, finance news, ...

Annotation: Given machine-translated snippet, answer the questions.

Quiz-Based Evaluation (2/2)

Moses 2007	Google 16.2.2010
Na provoz světla na roundabout, obrátit levice a projet ballymun. Otočit vlevo na křižovatce. ballymun / Collins Avenue Road Dcu je umístěna na Collins 500m na pravém boku Avenue.	Na semaforech na kruhový objezd, odbočit doleva a jet přes Ballymun. Odbočit vlevo na Collins Avenue / Ballymun silniční křižovatky. DCU se nachází na Collins Avenue 500 m na pravé straně.

Zaškrtněte pravdivá tvrzení:

1. DCU leží na Collins Avenue.
2. V daném městě mají na kruhových objezdech zřejmě semafory.
3. Při příjezdu budete mít DCU po levé straně.

Original: At the traffic lights on the roundabout, turn left and drive through Ballymun. Turn left at the Collins Avenue/Ballymun Road crossroads. DCU is located on Collins Avenue 500m on the right hand side.

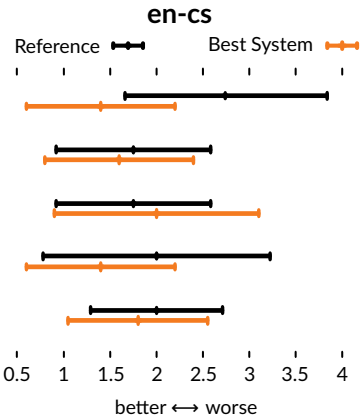
Correct answer: yyn

Maturita (GCSE)-Like (Vojtěchová et al., 2019)

Manual evaluation by domain experts, scoring in categories:

1. Language Resources – Spelling and Morphology
2. Vocabulary – Adequacy of Terms Used
3. Vocabulary – Clarity of the Text in Terms of Used Words
4. Syntax and Word Order
5. Coherence and Overall Understanding of the Text

plotted as average rank for better comparibility



Superhuman MT Translating Agreements?

Supplement No. 1 to the agreement on the sublease the apartment, of 13th May 2016 On the day, month and year written below Marta Burešová, pers. no. 695604/3017 Address: Radimova 8, Prague 6, 169 00 as the tenant on the one hand (Hereinafter referred to as "the tenant") and Karolína Černá, pers. no. 136205/891 Address: Alfrédova 13, Praha 4, 142 00 As a lessee on the other (Hereinafter referred to as "the lessee") collectively also referred to as "the Contracting parties" have agreed on this Supplement No. 1 to the Agreement on the sublease the apartment, of 13th May 2016 (hereinafter referred to as the "Supplement No. 1")

I. Introductory Provisions

On 13th May 2016, the tenant and the lessee closed the Agreement on the sublease of the apartment, under which the tenant let the lessee use the apartment No. 4 (area 49 m^2) of size 1+1/L in the ground floor of the house in Prague 4, Alfrédova 13, ...

(Vojtěchová et al., 2019)

Superhuman MT Translating Agreements?

Dodatek č. 1 ke smlouvě o podnájmu bytu ze dne 13. května 2016
V den, měsíc a rok níže napsané Marta Burešová, pers. no. 695604/3017 Adresa: Radimova 8, Praha 6, 169 00 jako nájemce na jedné straně (dále jen „nájemce“) a Karolína Černá, pers. no. 136205/891 Adresa: Alfrédova 13, Praha 4, 142 00 jako nájemce na straně druhé (dále jen „nájemce“) společně označované také jako „smluvní strany“ se dohodly na tomto dodatku č. 1 ke smlouvě o podnájmu, dále jen „nájemní smlouva“, dále jen „13. května 2016“).
I. Úvodní ustanovení
Dne 13. května 2016 nájemce a nájemce uzavřeli smlouvu o dalším pronájmu bytu, podle níž nájemce pronajímá nájemci byt č. 4 (plocha 49 m^2) o velikosti 1+1/l v přízemí domu v Praze 4, Alfrédova 13, ...

Superhuman MT Translating Agreements?

Dodatek č. 1 ke smlouvě o podnájmu bytu ze dne 13. května 2016
V den, měsíc a rok níže napsané Marta Burešová, pers. no. 695604/3017 Adresa: Radimova 8, Praha 6, 169 00 jako **nájemce na jedné straně** (dále jen „**nájemce**“) a Karolína Černá, pers. no. 136205/891 Adresa: Alfrédova 13, Praha 4, 142 00 jako **nájemce na straně druhé** (dále jen „**nájemce**“) společně označované také jako „smluvní strany“ se dohodly na tomto dodatku č. 1 ke smlouvě o podnájmu, dále jen „**nájemní smlouva**“, dále jen „13. května 2016“).
I. Úvodní ustanovení
Dne 13. května 2016 **nájemce a nájemce** uzavřeli smlouvu o dalším pronájmu bytu, podle níž **nájemce pronajímá nájemci** byt č. 4 (plocha 49 m^2) o velikosti 1+1/l v přízemí domu v Praze 4, Alfrédova 13

HMEANT (Lo and Wu, 2011)

- Improved evaluation of adequacy compared to BLEU.
- Reduced human labour of HTER (Snover et al., 2006).

Essence: Is the basic event structure understandable?

(*Who* did *what* to *whom*, *when*, *where* and *why*.)

1. Identify semantic frames and roles in ref & hyp.
   - Manual (5–15 min of training) or automatic (shallow SRL).
2. Mark match/partial/mismatch of each predicate and each argument.
   - Manual.
3. Calculate prec & rec across all frames in the sentence.
4. Report f-score.

HMEANT Illustration: Motivation

It is hard to rank A vs. B (even if we know R is the ref.)

Finally, he stood in the center of the referee Wolfgang Stark.

At the end of the day, was at the centre of a referee Wolfgang Stark.

The referee Wolfgang Stark then garnered some attention.

HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.

Finally, he stood in the center
of the referee Wolfgang Stark.

HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.

Finally, he stood in the center of the referee Wolfgang Stark.

Action

HMEANT Illustration: SRL

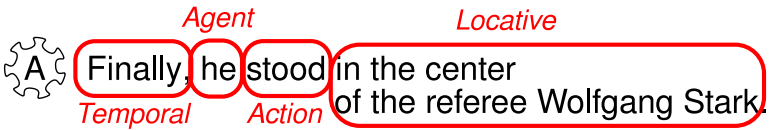
It is easier to mark roles of a single hypothesis.

Agent

Finally, he stood in the center
Action of the referee Wolfgang Stark.

HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.



HMEANT Illustration: SRL

The same SRL is performed on the reference.

The referee Wolfgang Stark then garnered some attention.

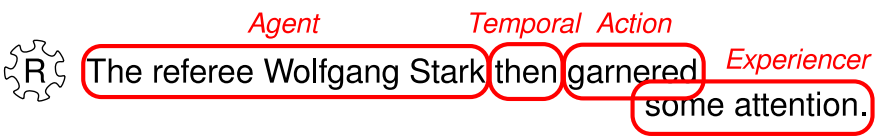
HMEANT Illustration: SRL

The same SRL is performed on the reference.

The referee Wolfgang Stark then *Action* garnered some attention.

HMEANT Illustration: SRL

The same SRL is performed on the reference.



HMEANT Illustration: Alignment

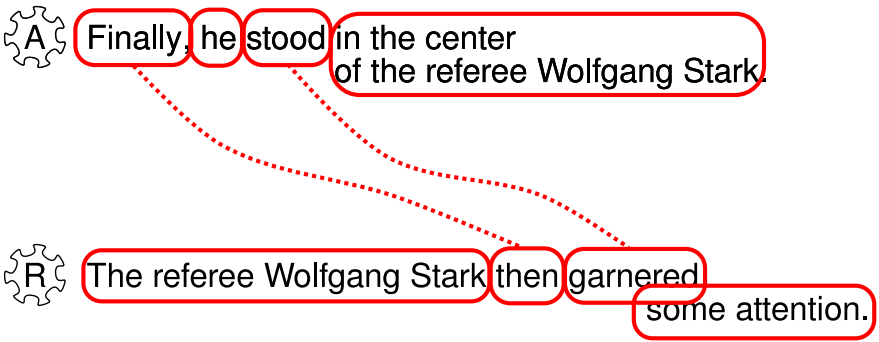
And finally, frames and role fillers are aligned.

Finally, he stood in the center of the referee Wolfgang Stark.

The referee Wolfgang Stark then garnered some attention.

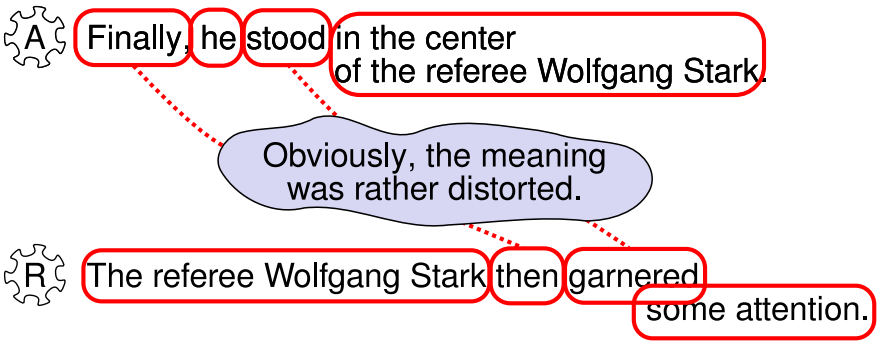
HMEANT Illustration: Alignment

And finally, frames and role fillers are aligned.



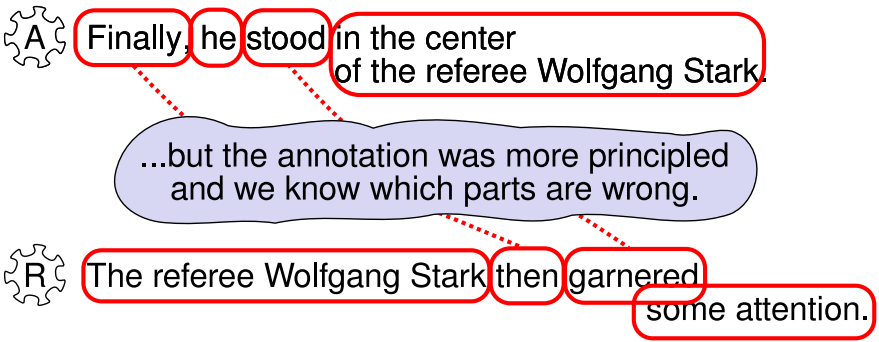
HMEANT Illustration

And finally, frames and role fillers are aligned.



HMEANT Illustration

And finally, frames and role fillers are aligned.



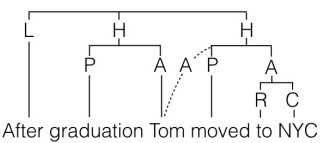
HUME

HUME (Birch et al., 2016) improves over HMEANT by:

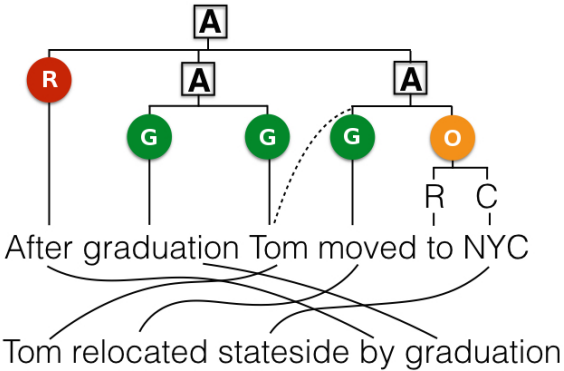
- using semantic trees (UCCA, Abend and Rappoport (2013)),
- using *source rather than reference*,
- using trees on the *source only*, not malformed hypothesis.

Two manual stages again:

1. Create UCCA tree for the source (can reuse for more systems!).
2. Label UCCA tree indicating how much was preserved by MT.

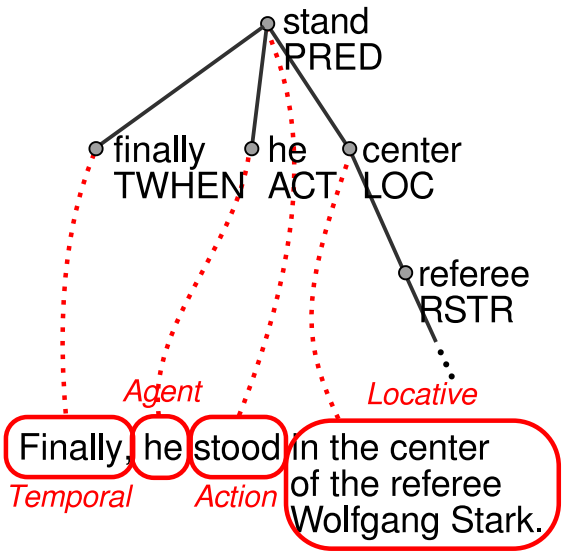


HUME Annotation



- Leafs get R/O/G (traffic lights): bad, mixed, good.
- Structure gets A/B: adequate, bad.

HMEANT/HUME are Close to FGD

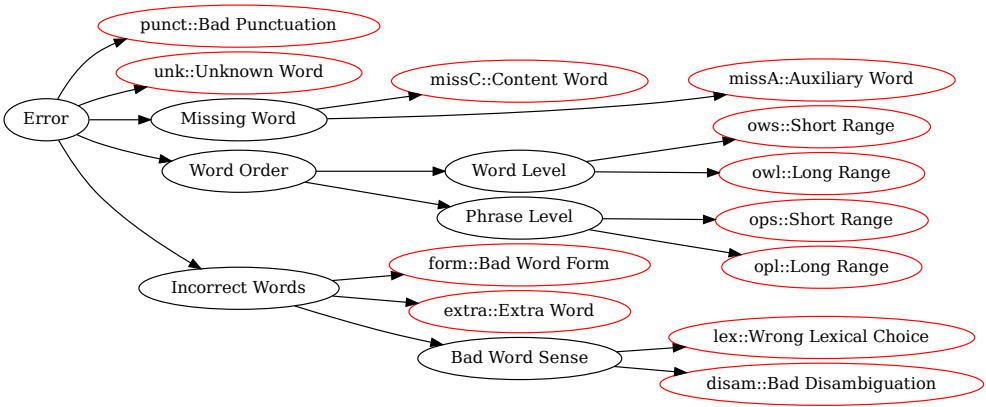


Project suggestion:
Use t-layer tools to:

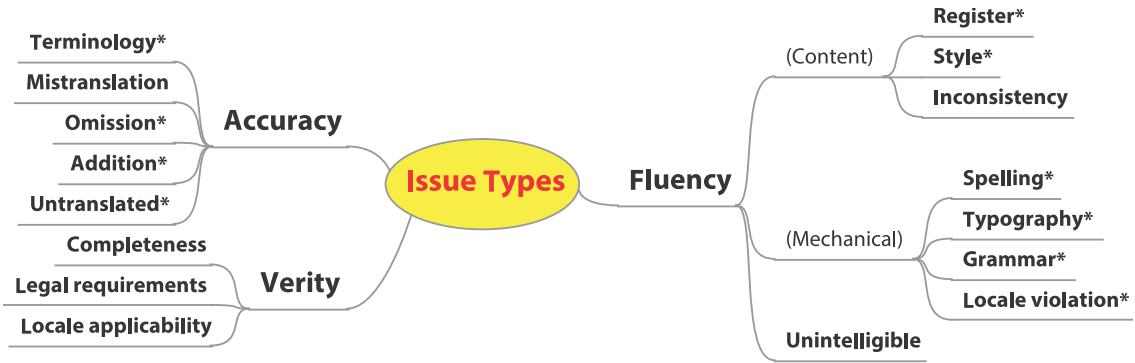
- Improve UCCA parser, or
- Automate: parse to UCCA or t-trees, predict R/O/G, A/B.

Evaluation by Flagging Errors

Classification of MT errors, following Vilar et al. (2006).

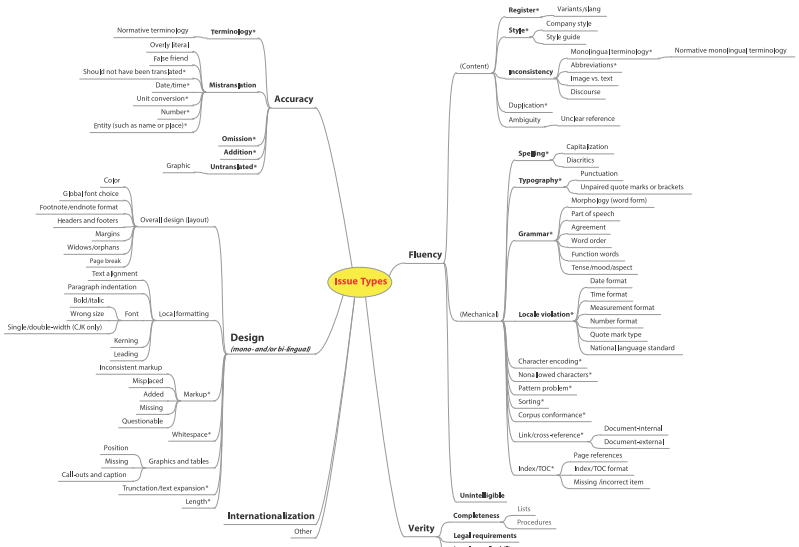


Standard MQM (Core)

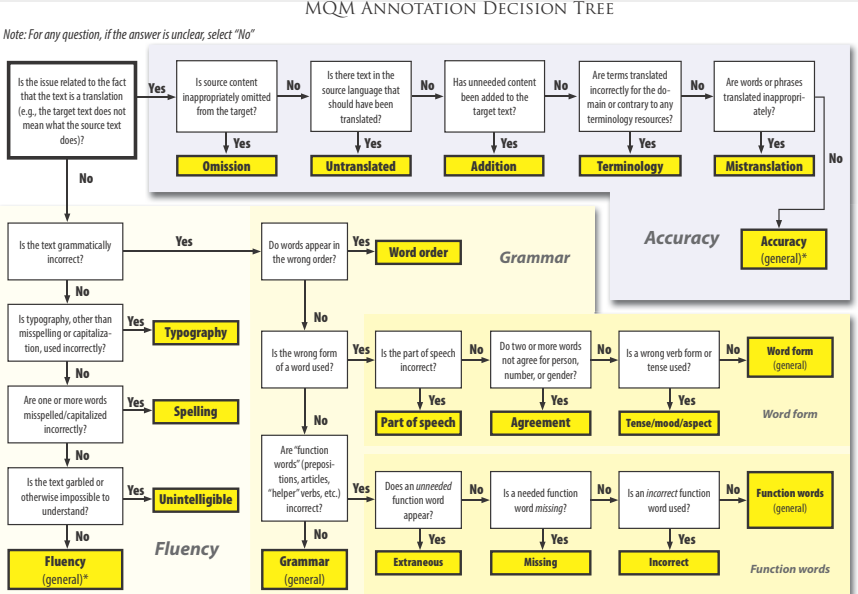


(Lommel et al., 2014)

Standard MQM (Overkill)



MQM Decision Tree (Simplified)

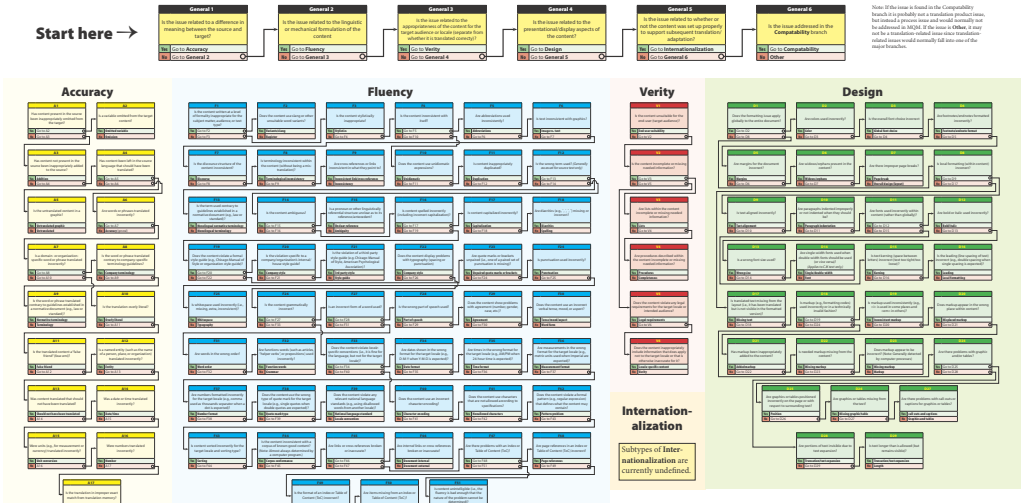


MQM Decision Tree (Full)

Multidimensional Quality Metrics (MQM): Full Decision Tree

The Multidimensional Quality Metrics (MQM) Framework provides a hierarchical categorization of error types that occur in translated or localized products. Based on a detailed analysis of existing translation quality metrics, it provides a flexible typology of *issue types* that can be applied to analytic or holistic translation quality evaluation tasks. Although the full MQM issue tree (which, as of November 2014, contains 115 issue types categorized into five major branches) is not intended to be used in its entirety for any particular evaluation task, this overview chart presents a "decision tree" suitable for selecting an issue type from it. In practical terms, however, an individual metric would have a smaller decision tree that covers just the issues contained in that metric.

To use the decision tree start with the first question and follow the appropriate answers until a specific issue type is reached.



Error Flagging Example

Annotation rules:

- Mark/suggest as little as necessary.
- Compare to *source*, not to *reference*. Literal translation ok.
- Preserve white space. Don't add or remove word/line breaks.
- Only insert error labels followed by : :.
- For missing words, use _ instead of space, if necessary.

Src	Perhaps there are better times ahead.
Ref	Možná se tedy blýská na lepší časy.
	Možná, že extra::tam jsou lepší disam::krát lex::dopředu.
	Možná extra::tam jsou příhodnější časy vpředu.
missC::v_budoucnu	Možná form::je lepší časy.
	Možná jsou lepší časy lex::vpřed.

Results on WMT09 Dataset

	google	cu-bojar	pctrans	cu-tectomt	Total
Automatic: BLEU	13.59	14.24	9.42	7.29	–
Manual: Rank	0.66	0.61	0.67	0.48	–
disam	406	379	569	659	2013
lex	211	208	231	340	990
Total bad word sense	617	587	800	999	3003
missA	84	111	96	138	429
missC	72	199	42	108	421
Total missed words	156	310	138	246	850
form	783	735	762	713	2993
extra	381	313	353	394	1441
unk	51	53	56	97	257
Total serious errors	1988	1998	2109	2449	8544
ows	117	100	157	155	529
punct	115	117	150	192	574
...	...	...	...	...	...
tokenization	7	12	10	6	35
Total errors	2319	2354	2536	2895	10104

Contradictions in Manual Evaluation

Results for WMT10:

Evaluation Method	Google	CU-Bojar	PC Translator	TectoMT
$\geq$ others (WMT10 official)	**70.4**	65.6	62.1	60.1
$>$ others	49.1	45.0	**49.4**	44.1
Edits deemed acceptable [%]	**55**	40	43	34
Quiz-based evaluation [%]	80.3	75.9	80.0	**81.5**
Automatic: BLEU	**0.16**	0.15	0.10	0.12
Automatic: NIST	**5.46**	5.30	4.44	5.10

Results for WMT19:

- Best systems match humans in GCSE-like scoring.
- They score worse in pseudo-doc-aware DA.
- They are absolutely terrible on agreements.

... each technique provides a different picture.

Problems of Manual Evaluation

- Expensive in terms of time/money.
- Subjective (some judges are more careful/better at guessing).
- Not quite consistent judgments from different people.
- Not quite consistent judgments from a single person!
- Not reproducible (too easy to solve a task for the second time).
- Experiment design is critical!

- Black-box evaluation important for users/sponsors.
- Gray/Glass-box evaluation important for the developers.
- SRC-based allows to compare with humans.
- Sentence-level no longer relevant for large language pairs.

Automatic Evaluation

- Comparing MT output to reference translation.
- (Reference-less evaluation is called Quality Estimation.)

- Fast and cheap.
- Deterministic, replicable.
- Allows automatic model optimization ("tuning", MERT).

- Usually good for checking progress.
- Usually bad for comparing systems of different types.

BLEU (Papineni et al., 2002)

- Based on geometric mean of n-gram precision.

$\approx$ ratio of 1- to 4-grams of hypothesis confirmed by a ref. translation

Src	The legislators hope that it will be approved in the next few days .	Confirmed
Ref	Zákonodárci doufají , že bude schválen v příštích několika dnech .	1 2 3 4
Moses	**Zákonodárci doufají , že bude schválen v** nejbližších dnech .	9 7 5 4
TectoMT	**Zákonodárci doufají , že bude** schváleno další páru volna .	6 4 3 2
Google	Zákonodárci naději **, že bude schválen v** několika příštích dnů .	9 4 3 2
PC Tr.	Zákonodárci doufají že to bude schválený v nejbližších dnech .	7 2 0 0

n-grams confirmed: none, unigram, bigram, *trigram*, **fourgram**

E.g. Moses produced 10 unigrams (9 confirmed), 9 bigrams (7 confirmed), ...

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{4}\log\left(\frac{9}{10}\right) + \frac{1}{4}\log\left(\frac{7}{9}\right) + \frac{1}{4}\log\left(\frac{5}{8}\right) + \frac{1}{4}\log\left(\frac{4}{7}\right)\right)$$

BP is "brevity penalty"; $\frac{1}{4}$ are uniform weights, the "denominator" equivalent for $\sqrt[4]{\cdot}$ in geometric mean in the log domain.

BLEU: Avoid Cheating/Gaming the Metric

- Confirmed counts "clipped" to avoid overgeneration.
- "Brevity penalty" applied to avoid too short output:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

Ref 1:	The cat is on the mat .
Ref 2:	There is a cat on the mat .
Candidate:	The the the the the the the .
	$\Rightarrow$ Clipping: only $\frac{3}{8}$ unigrams confirmed.
Candidate:	The the .
	$\Rightarrow$ $\frac{3}{3}$ unigrams confirmed but the output is too short.
	$\Rightarrow$ BP $= e^{1-7/3} = 0.26$ strikes.

The candidate length c and "effective" ref. length r calculated over the whole test set.

BLEU Properties

- Within the range 0-1, often written as 0 to 100%.
- Human translation against other humans: ~60%
- Google Chinese→English: ~30%, Arabic→English: ~50%.

- BLEU for individual sentences not reliable.
- More so with only 1 reference translation:

Src	" We ' ve made great progress .
Ref	" Učinili jsme velký pokrok .
Moses	" my jsme udělali *velký pokrok .*
TectoMT	" Udělali **jsme velký pokrok .**
Google	" My jsme dosáhli obrovského pokroku .
PC Translator	" udělali **jsme velký pokrok .**

Test Set Influence on BLEU

Havlíček (2007) evaluates the influence of:

- number of reference translations,
- translation direction.

on *human-produced* text (1 human translation against 4 others).

	cs→en, Professionals				en→cs, Math Students			
Refs	Indiv. Results			**Avg**	Indiv. Results			**Avg**
1	41.15	32.66	34.03	**35.95**	3.66	8.62	5.79	**6.02**
2	49.09	49.78	41.26	**46.71**	9.82	8.26	9.36	**9.15**
3	52.63			**52.63**	13.06			**13.06**

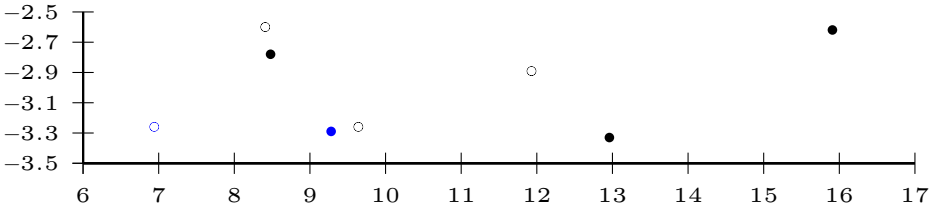
⇒ heavy dependence on the number of references.

More references allow to match more n-grams of MT output.

⇒ heavy dependence on the translation direction and quality.

Correlation with Human Judgments

BLEU scores vs. human rank, the higher, the better:



WMT08 Results	In-domain •		Out-of-domain ◦	
	BLEU	Rank	BLEU	Rank
Factored Moses	**15.91**	**-2.62**	**11.93**	-2.89
PC Translator	8.48	-2.78	8.41 !!	**-2.60**
TectoMT	9.28	-3.29	6.94	-3.26
Vanilla Moses	12.96	-3.33	9.64	-3.26

⇒ PC Translator nearly won Rank but nearly lost in BLEU.

Dirty Tricks

- PCEDT 1.0 (Čmejrek et al., 2004) contains test set with:
  - 1 English original,
  - 1 Czech translation,
  - 4 English back-translations (via Czech).
- Čmejrek et al. (2003) evaluate cs→en MT using all 5 English sentences: they include the original source among the references and report 5-fold average of BLEU (on 4 refs).
- The additional accepted variance in output increases BLEU compared to BLEU on the 4 back-translations only.

	5-fold Avg of 4-BLEU	4 refs only
PBT, no additional LM	34.8±1.3	32.5
PBT, bigger LM	36.4±1.3	34.2
PBT, more parallel texts, bigger LM	38.1±0.8	36.8

Improving BLEU in cs→en MT

A summary of older experiments. (Bojar et al., 2006; Bojar, 2006)

Deterministic pre- and post-processing	
similar tokenization of reference	+10.0 !!!
lemmatization for alignment	+2.0
handling numbers	+0.9
fixing clear BLEU errors	+0.5 !
dependency-based corpus expansion	+0.3
More parallel or target-side monolingual data	
out-of-domain parallel texts, bigger in-domain LM	+5.0
bigged in-domain LM	+1.7
out-of-domain parallel texts, also in LM	+0.4
adding a raw dictionary	+0.2

- Complicated methods bring a little.
- Data bring more.
- Huge jumps from superficial properties but just higher BLEU, same MT quality.

Finding Clear BLEU Losses

Missing bigram = all references contained it but not the hypothesis.
Superfluous bigram = the hypothesis contained it but none of the references.

Top missing bigrams:

19	, "	12	" said
12	of the	10	Free Europe
10	Radio Free	7	. "
6	L.J. Hooker	6	United States
6	in the	6	the United
6	the strike		...

Top superfluous bigrams:

26	, ''	18	'' .
14	" said	12	, which
11	Svobodná Evropa	8	, when
8	the state	7	, who
7	J. Hooker	7	L. J.
7	company GM		...

Four simple rules to improve BLEU by +0.2 to +0.5 on a particular test set:

'' .	→	. "	L. J. Hooker	→	L.J. Hooker
''	→	"	the U.S.	→	the United States

Technical Problems of BLEU

BLEU scores are not comparable:

- across languages.
- on different test sets.
- with different number of reference translations.
- with different implementations of the evaluation tool.
- There are different definitions of "reference length":
  Papineni et al. (2002) not specific. One can choose the shortest, longest, average, closest (the smaller or the larger!).
- Very sensitive to tokenization:
  Beware esp. of malformed tokenization of Czech by foreign tools.

$\Rightarrow$ Use a fixed implementation, e.g. sacreBLEU (Post, 2018).

Fundamenal Problems of BLEU

- BLEU overly sensitive to word forms and sequences of tokens.

Confirmed by Ref	Contains Error Flags	1-grams	2-grams	3-grams	4-grams
Yes	Yes	6.34%	1.58%	0.55%	0.29%
Yes	No	36.93%	13.68%	5.87%	2.69%
No	Yes	22.33%	41.83%	54.64%	63.88%
No	No	**34.40%**	**42.91%**	**38.94%**	**33.14%**
Total n-grams		35 531	33 891	32 251	30 611

30–40% of tokens not confirmed by reference but without errors.
⇒ Enough space for MT systems to differ unnoticed.
⇒ Low BLEU scores correlate even less.

Fixing Fundamenal Issues of BLEU

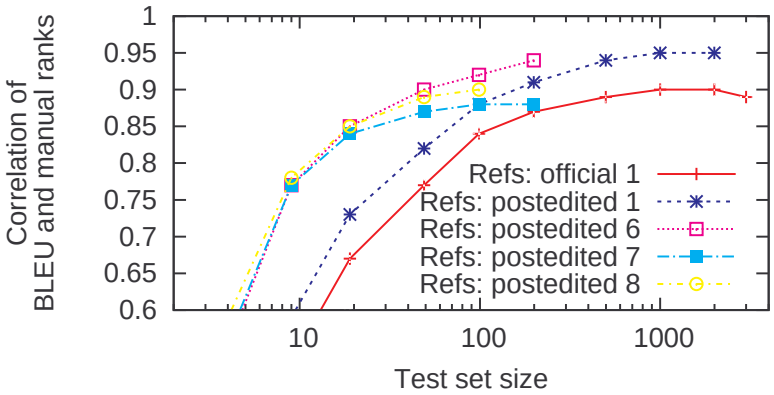
Evaluate coarser units:

- Lemmas or deep-lemmas instead of word forms:
  - e.g. SemPOS (Kos and Bojar, 2009): bags of t-lemmas.
- Sequences of characters:
  - e.g. chrF3 (Popović, 2015): F-score of character 6-grams.
- Use shorter of gappy sequences:
  - e.g. BEER (Stanojevic and Sima'an, 2014) uses characters and also pairs of (not necessarily adjacent) words.

Use better references:

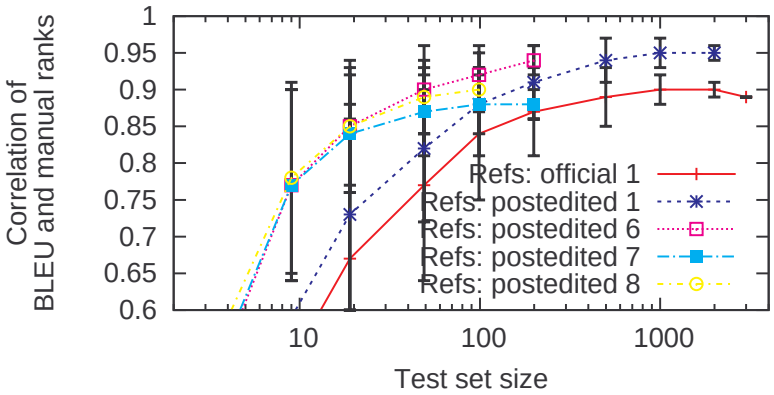
- Using more references alone helps.
- Post-edited references serve better.
  - e.g. HTER (Snover et al., 2006): Measuring edit distance to manually corrected output.

Post-Edited References Serve Better



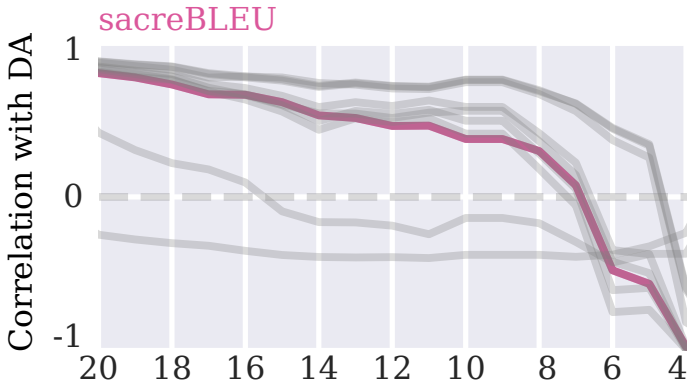
- Refs created by post-editing serve better than independent ones.
- 100 sents with 6–7 postedited refs as good as 3k indep refs.

Post-Edited Refs Better



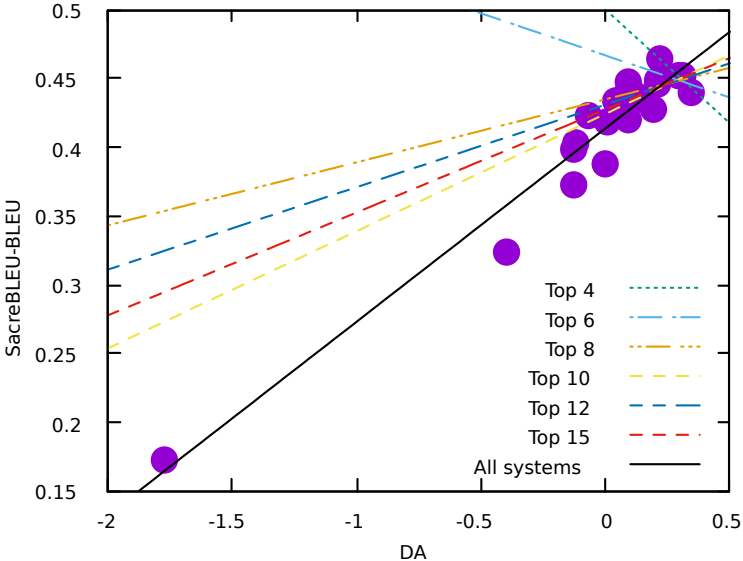
- ... but error bars quite wide
  $\Rightarrow$ specific sentences important.

Fundamenal Problem of Correlation



- Correlation depends on the *underlying set of MT systems*.
- Often poor correlation when only top-scoring systems are considered, see Ma et al. (2019).

Fundamenal Problem of Correlation



Empirical Confidence Intervals

In statistics, confidence intervals indicate how well was a parameter (e.g. the mean) of a random variable with *known/assumed distribution* estimated from a set of repeated measurements.

- We don't want to assume any distribution!
- How to "repeat" experiments with a deterministic MT system?

Use "bootstrapping" (Koehn, 2004):

1. Obtain 1000 different test sets:
   Randomly select sents., repeat some, ignore some, preserving test set size.
2. Sort by the score.
3. Drop top and bottom 2.5% (i.e. 25 out of 1000) results.

$\Rightarrow$ The lowest and highest remaining scores are 95% empirical confidence interval around the score obtained on the full test set.

End-to-end vs. Component Eval.

- Similar to black vs. glass box evaluation and translation vs. task-based evaluation.
- Evaluation of a single component may not correlate with overall performance of the system.

Pre-processing	Symmetrization	Alignment Error Rate	BLEU
Lemmas + singletons	Intersection	**14.6**	30.8
Lemmas	Intersection	15.0	29.8
Lemmas	Union	17.2	**32.0**
Lemmas + singletons	Union	17.4	31.9
Baseline (word forms)	Union	25.5	29.8
Baseline (word forms)	Intersection	27.4	28.2

Data by Bojar et al. (2006). See also e.g. Lopez and Resnik (2006).

Summary, The Moral of the Story

Metrics drive research:

- Measure the property that "saves money" in your application.
- Design automatic metrics to correlate with humans.

Comparisons of automatic scores trustworthy
only under all the following:

- a single test set was used (of your domain of interest),
- evaluated by a single evaluation tool (hopefully without bugs),
  E.g. for BLEU different tools tokenize and define ref. length differently.
- the metric reflects your final objective (AER vs. BLEU),
- confidence intervals are estimated.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria, August. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-Based Evaluation of Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas, November. Association for Computational Linguistics. peer-reviewed.

Hervé Blanchon, Christian Boitet, and Laurent Besacier. 2004. Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals. In *Proceedings of International Conference on Spoken Language Processing ICSLP 2004*, Jeju Island, Korea, October.

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August. Springer.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium, October. Association for