

Fine-grained Human Evaluation of Neural versus Phrase-based Machine Translation

EAMT, Praha, 31st May 2017

Filip Klubička

Antonio Toral

Víctor M. Sánchez-Cartagena

University of Zagreb

University of Groningen

Prompsit Language Engineering

Introduction

In many setups, NMT has surpassed the performance of the mainstream MT approach to date: PBMT

In many setups, NMT has surpassed the performance of the mainstream MT approach to date: PBMT

E.g. news translation shared task at WMT'16

- 10 language directions: EN \leftrightarrow CS, DE, FI, RO, RU
- Automatic evaluation: BLEU, TER
- Human evaluation: ranking translations

Overall Evaluation (Automatic)

System	CS	DE	FI	RO	RU
	From EN				
PBMT	23.7	30.6	15.3	27.4	24.3
NMT	25.9	34.2	18.0	28.9	26.0
	Into EN				
PBMT	30.4	35.2	23.7	35.4	29.3
NMT	31.4	38.7	-	34.1	28.2

Table 1: BLEU scores of the best NMT and PBMT systems

Bold: statistical significance

Overall Evaluation (human)

System	CS	DE	FI	RO	RU
	From EN				
PBMT	23.7	30.6	15.3	27.4	24.3
NMT	25.9	34.2	18.0	28.9	26.0
	Into EN				
PBMT	30.4	35.2	23.7	35.4	29.3
NMT	31.4	38.7	-	34.1	28.2

Table 2: BLEU scores of the best NMT and PBMT systems

Bold: statistical significance (BLEU)

Green: statistical significance (human evaluation)

Overall, NMT outperforms PBMT, but... which are its strengths? And what are its weaknesses?



Paper	Direction	Findings: NMT...
Bentivogli et al., 2016	EN→DE	<ol style="list-style-type: none"><li data-bbox="786 234 1215 317">1. Improves on reordering and inflection<li data-bbox="786 337 1140 368">2. Decreases PE effort<li data-bbox="786 389 1215 472">3. Degrades with sentence length

Background

Paper	Direction	Findings: NMT...
Bentivogli et al., 2016	EN→DE	<ol style="list-style-type: none">1. Improves on reordering and inflection2. Decreases PE effort3. Degrades with sentence length
Toral and Sánchez-Cartagena, 2017	EN→CS, DE, FI, RO, RU CS, DE, RO, RU→EN	<ol style="list-style-type: none">1. Corroborated findings 1 and 2 from Bentivogli2. Higher inter-system variability3. More reordering than PBMT but less than hierarchical PBMT

Limitations of these analyses

- **Performed automatically.** E.g. inflection errors detected with a PoS tagger
- **Coarse-grained.** 3 error types: inflection, reordering and lexical

Limitations of these analyses

- **Performed automatically.** E.g. inflection errors detected with a PoS tagger
- **Coarse-grained.** 3 error types: inflection, reordering and lexical



This work: fine-grained human analysis of NMT vs PBMT and factored PBMT

- **Fine-grained.** Errors annotated following a detailed error taxonomy (>20 error types)
- **Human.** Errors annotated manually
- **Factored PBMT.** Not compared to NMT to date¹
- **Direction.** English-to-Croatian, i.e. MT into a morphologically-rich target language, challenge for phenomena such as agreement (case, gender, number)

¹To the best of our knowledge

Data sets and MT systems

- **Dev.** First 1k sentences from English test set at WMT'12, translated into Croatian
- **Test.** Same but from WMT'13
- **Train**
 - **Parallel.** 4.8M sentence pairs selected according to cross-entropy from different sources: EU/legal, news, web, subtitles
 - **Monolingual.** Web + target side of parallel data

All systems trained on the same data set. NMT does **not** use monolingual data.

- **Pure PBMT**. Standard Moses + hierarchical reordering, bilingual neural LM, OSM
- **Factored PBMT**. Maps 1 factor in the source (surface form) to 2 in the target (surface form and morphosyntactic description)
- **NMT**
 - Sequence-to-sequence with attention
 - Unsupervised word segmentation (byte pair encoding)
 - Trained for 10 days, models saved every 4.5h. Ensemble of 4 best models on dev set

Results with automatic metrics

System	BLEU	TER
PBMT	0.2544	0.6081
Factored PBMT	0.2700	0.5963
NMT	0.3085	0.5552

Human Evaluation

Multidimensional Quality Metrics (MQM)

- Framework for defining custom quality metrics
- Provides a flexible vocabulary of quality issue types

Multidimensional Quality Metrics (MQM)

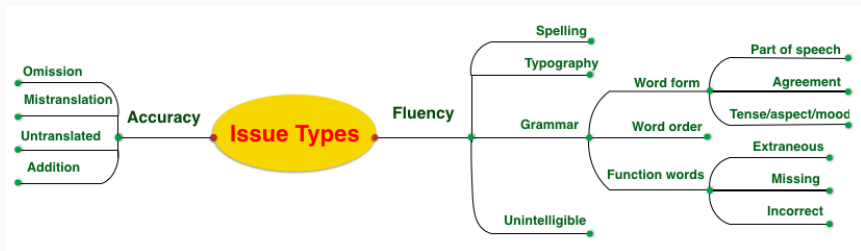
- Framework for defining custom quality metrics
- Provides a flexible vocabulary of quality issue types

We devised an MQM-compliant taxonomy with these aims

- Right level of **granularity**: trade-off between having a detailed taxonomy and the annotation process being viable
- Error types relevant for the **translation direction**

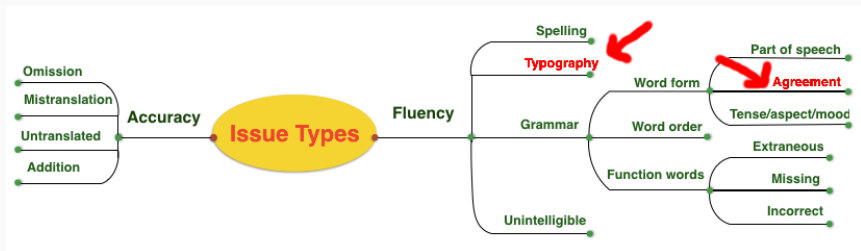
Error taxonomy

MQM core taxonomy



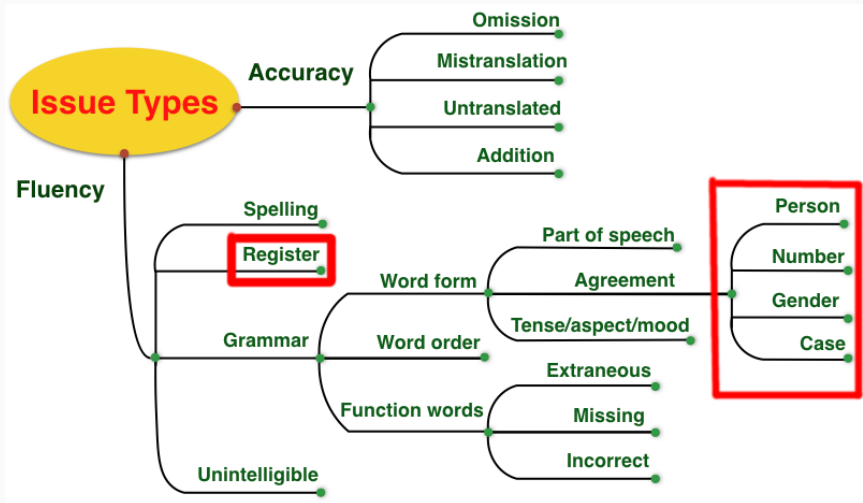
Error taxonomy

MQM core taxonomy



Error taxonomy

MQM Slavic taxonomy



Annotation Setup

- Tool: `translate5`
- 2 annotators (native Croatian, C1 English)
- 100 randomly selected sentences from the test set annotated
 - Total: 600 annotated sentences (100 sentences * 3 systems * 2 annotators)

Annotation Process



Authenticated user: Filip Kubicka
Login name: fkubicka
Task: m3-fm_enhr_case_rnd

Logout Tasks My settings English

Segment list and editor

Editor modes: Hide tags Short tag view Full tag view Reset sorting / filtering QM Subsegment Statistics

Nr.	source	reference	mt_out1	mt_out2	Comments
1	A shuttle bus brings people to it from the foot of the mountain.	Direktni bus koji kreće s podnožja planine prevozí ljude do vrha.	18 Shuttle bus 21 dovodi ljude 22 na to 23. 192 iz 221 podnožja planine.	18 Shuttle bus 21 dovodi ljude 22 na to 23. 192 iz 221 podnožja planine.	
2	Now, everybody is divided into micro societies, it's hard to be liked by everyone.	Sada su svi podijeljeni u mikrodruštva, teško je sviđati se svima.	Sada, svi se 114 dijele 144 na mikro 116 društava 164, teško je 17 da se sviđimo 27 drugima.	Sada, svi su podijeljeni u mikro društva, teško je 18 da se sviđimo 28 drugima.	mt_out1 phrase agr Filip Kubicka 2016-09-09 ... [1 more comment]
3	Right now, even if I need a few knuckledusters, I get them through someone I trust.	Trenutno, ako trebam par boksera, nabavit ću ih preko nekog kome vjerujem.	Sada, čak i ako trebam nekoliko 15 knuckledusters 25, dobivam ih kroz 18 nekog 181 iz 19 vjerujem.	Sada, čak i ako trebam nekoliko 15 knuckledusters 25, dobivam ih kroz nekoga 18 iz 19 vjerujem.	mt_out1 phrase agr Filip Kubicka 2016-09-09
4	Once this number has been found, the other bones can easily check that it is the right one.	Jednom kad je taj broj pronađen, ostali čvorovi lako mogu provjeriti radi li se o pravoj transakciji.	Kad ovaj broj 116 > 111 pronađen, drugi čvorovi lako mogu provjeriti da je ispravna.	Nakon što je ovaj broj 114 < 111 pronađen, ostali čvorovi mogu lako provjeriti da je ispravna.	
5	Because, in the opinion of a number of people working in palliative care, great moments occur at the very heart of such regression.	Jer, prema mišljenju ljudi koji rade u palijativnoj skrbi, veliki trenuci dešavaju se u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u 116 palijativnoj 120 skrbi, veliki trenuci se 127 odvijaju 27 u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u palijativnoj skrbi, veliki trenuci se javljaju u samom srcu takve regresije.	mt_out1 phrase agr Filip Kubicka 2016-09-09
6	But we'll have something to say against some very good European teams.	Ali imat ćemo nešto za reći protiv nekih vrlo dobrih europskih timova.	Ali ćemo imati nešto za reći protiv 12 neke vrlo dobre europske 21 momčadi.	Ali ćemo imati nešto za reći protiv vrlo dobrih europskih timova.	
7	The Ministry of the Interior does not put arms from the illegal market back into circulation	Ministarstvo unutarnjih poslova ne vraća oružje s crnog tržišta natrag u cirkulaciju	Ministarstvo unutarnjih poslova ne stavi 12 ruke 21 na 221 ilegalnom tržištu natrag u promet	Ministarstvo unutarnjih poslova ne stavi 12 ruke 21 od nelegalnog tržišta natrag u promet	
8	People used to get together in flocks, Bohemians liked one thing, the simple people, something else.	Ljudi su se prije nalazili u jatima, boemima se sviđala jedna stvar, jednostavni ljudi, nešto drugo.	Ljudi 12 koriste 21 da se okupe u jatima, 12 primilo 21 je volio jednu stvar, jednostavni ljudi, nešto drugo.	Ljudi 12 koji se koriste 21, kako bi se 12 zajedno 21 u jatima, 4 gornja 41 boema sviđala jedna stvar, jednostavni ljudi, nešto drugo.	
9	How do you explain this progression?	Kako objašnjavate taj napredak?	Kako 12 objašnjavati 21 ovo 12 napredovanje 31?	Kako 12 objašnjavati 21 ovo 12 napredovanje 21?	
10	If a migrant does not understand the language, says Sebelev with certainty, he is doomed to come across unconscious people, who, pretending to help, will force upon him a "ticket" to terrible. cramed	"Ako migrant ne razumije jezik", sa sigurnošću kaže Sebelev, "sudeći mu je susret s nesavjesnim ljudima koji će mu, praveći se da pomažu, dati "kartu" za crnoze. sretnosne barake u kojima će novci	Ako 112 još ne 114 razumiju 114 jezik, 118 kaže 21 da 21 sa sigurnošću Sebelev 118, on je osuđen 21 211 da nesavjesni ljudi, koji, prevarajući se da 12 114	Ako dosejnik ne 114 razumiju 21 jezik, 118 kaže Sebelev sa sigurnošću 118, on je osuđen da 21 se preko 21 nesavjesni ljudi, koji 12 su 41, prevarajući se 12 u pomoć	mt_out2 sentence Filip Kubicka 2016-09-09 ... [3 more comments]

Segment meta data

- Terminology: Keine Terminologie vorhanden!
- QM Subsegments: add QM Subsegment
- Severity: Critical
- Comment:
- QM: OK Minor errors Must be reworked
- Status: Status 1 Status 2 Status 3 Nicht gesetzt

Comments for the current segment

Annotation Process

translate5

Authenticated user: Filip Kubicka
Login name: fkubicka
Task: m3-fm_enhr_case_rnd

Logout Tasks My settings English

Segment list and editor

Editor modes: Hide tags Short tag view Full tag view Reset sorting / filtering QM Subsegment Statistics

Nr.	source	reference	mt_out1	mt_out2	Comments
1	A shuttle bus brings people to it from the foot of the mountain.	Direktni bus koji kreće s podnožja planine prevozi ljude do vrha.	[5 Shuttle bus 5] dovodi ljude [2 na to 2] [22 iz 22] podnožja planine.	[5 Shuttle bus 5] dovodi ljude [2 na to 2] [22 iz 22] podnožja planine.	
2	Now, everybody is divided into micro societies, it's hard to be liked by everyone.	Sada su svi podjeljeni u mikrodružva, teško je sviđati se svima.	Sada, svi se [11 dije] [14] na mikro [16] društava [16], teško je [7 da se sviđimo 7] drugima.	Sada, svi su podjeljeni u mikro društva, teško je [11 da se sviđimo 21] drugima.	mt_out1 phrase as Filip Kubicka 2016-09-09 ... [1 more comment]
3	reference	reference	mt_out1	mt_out2	mt_out1 phrase as Filip Kubicka 2016-09-09
4	Direktni bus koji kreće s podnožja planine prevozi ljude do vrha.	[5 Shuttle bus 5] dovodi ljude [2 na to 2] [22 iz 22] podnožja planine.			
5	Because, in the opinion of a number of people working in palliative care, great moments occur at the very heart of such regression.	Jer, prema mišljenju ljudi koji rade u palijativnoj skrbi, veliki trenuci dešavaju se u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u [11 palijativne 16] skrbi, veliki trenuci se [11 odvijaju 11] u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u palijativnoj skrbi, veliki trenuci se javljaju u samom srcu takve regresije.	mt_out1 phrase as Filip Kubicka 2016-09-09
6	But we'll have something to say against some very good European teams.	Ali imat ćemo nešto za reći protiv nekih vrlo dobrih europskih timova.	Ali ćemo imati nešto za reći protiv [11] neke vrlo dobre europske [21] momčadi.	Ali ćemo imati nešto za reći protiv vrlo dobrih europskih timova.	
7	The Ministry of the Interior does not put arms from the illegal market back into circulation	Ministarstvo unutarnjih poslova ne vraća oružje s crnog tržišta natrag u cirkulaciju	Ministarstvo unutarnjih poslova ne stavi [11] ruke [21] [22 na 22] ilegalnom tržištu natrag u promet	Ministarstvo unutarnjih poslova ne stavi [11] ruke [21] od nelegalnog tržišta natrag u promet	
8	People used to get together in flocks, Bohemians liked one thing, the simple people, something else.	Ljudi su se prije nalazili u jatima, boemima se sviđala jedna stvar, jednostavni ljudi, nešto drugačije.	Ljudi [12 koriste 21] da se okupe u jatima, [12 primio 21] je volio jednu stvar, jednostavni ljudi, nešto drugo.	Ljudi [12 koji se koriste 21] kako bi se [12 zajedno 21] u jatima, [14 gomila 4] boema sviđala jedna stvar, jednostavni ljudi, nešto drugo.	
9	How do you explain this progression?	Kako objašnjavate taj napredak?	Kako [11 objašnjavati 21] ovo [12] napredovanje [11]?	Kako [11 objašnjavati 21] ovo [12] napredovanje [11]?	
10	If a migrant does not understand the language, says Sebelev with certainty, he is doomed to come across unconscious people, who, pretending to help, will force upon him a "ticket" to terrible. cramed	"Ako migrant ne razumije jezik", sa sigurnošću kaže Sebelev, "sudeći mu je susret s nesavjesnim ljudima koji će mu, praveći se da pomažu, dati "kartu" za crno. nesvesne barake u kojima će novci	Ako [11] [11] [11] ne [11] razumiju [11] jezik, [11] kaže [11] da [11] sa sigurnošću Sebelev [11], on je osuđen [11] [11] da nesavjesni ljudi, koji, prevarajući se da [11] [11]	Ako dosejnik ne [11] razumiju [11] jezik, [11] kaže Sebelev sa sigurnošću [11], on je osuđen da [11] se preko [11] nesavjesni ljudi, koji [11] su [11], prevarajući se [11] u pomoć	mt_out2 sentence Filip Kubicka 2016-09-09 ... [3 more comments]

Segment meta data

- Terminology: Keine Terminologie vorhanden!
- QM Subsegments: add QM Subsegment
- Severity: Critical
- Comment:
- QM: OK, Minor errors, Must be reworked
- Status: Status 1, Status 2, Status 3, Nicht gesetzt

Comments for the current segment

Annotation Process

u podijeljeni u mikro društva, [7] da se svidimo [7] drugima.

mt_out1 frase agr
Filip Klubicka 2016-09-0
... (1 more comment)

i ako trebam nekoliko [5]
sters [5], dobivam ih kroz nekoga
rujem.

je ova
ovi mo

iljenju
toj skr
arcu takve regresije.

nati nešto za reći protiv vrlo dobrih
imova.

vo unutarnjih poslova ne stavi [2]

Register [7]
Spelling [8]
Grammar [9]
Unintelligible [23]

Word form [10]
Word order [18]
Function words [19]

mt_out1 phrase agr
Filip Klubicka 2016-09-0

QM Subsegments
add QM Subsegment ▾
Accuracy [1]
Fluency [6]

QM
 OK
 Minor errors
 Must be reworked

Status
 Status 1
 Status 2
 Status 3
 Nicht gesetzt

Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter annotator agreement for each MT system

PBMT	Factored	NMT	Concatenated
0.56	0.49	0.44	0.51

Inter Annotator Agreement

Calculated at sentence level with Cohen's κ

Inter annotator agreement for each MT system

PBMT	Factored	NMT	Concatenated
0.56	0.49	0.44	0.51

Inter annotator agreement for each error type (min: 0.27, max: 0.72)

Notes

- **Outputs** have different length
 - normalise errors by number of tokens: ratio of tokens with and without errors
- **Statistical significance** with χ^2
 - 2x2 contingency tables for each pair of systems: (PBMT, factored), (PBMT, NMT), (factored, NMT)
- **Error types**: concatenated and separately

Results

Overall: considering all error types

	PBMT		Factored		NMT	
	No error	Error	No error	Error	No error	Error
Overall	2826	1010	3007	**809	3199	**469

** $p < 0.01$ (compared to the system on its left)

Results

Overall: considering all error types

	PBMT		Factored		NMT	
	No error	Error	No error	Error	No error	Error
Overall	2826	1010	3007	**809	3199	**469

** $p < 0.01$ (compared to the system on its left)

Relative reduction of errors:

- Factored: 20%
- NMT: 42% (wrt factored), 54% (wrt PBMT)

Results (by error type, accuracy branch)

Error type	PBMT		Factored		NMT	
	No error	Error	No error	Error	No error	Error
Accuracy	3467	369	3525	*291	3402	266
Mistranslation	3547	289	3586	*230	3471	197
Omission	3801	35	3793	23	3619	*49
Addition	3814	22	3797	19	3655	13
Untranslated	3813	23	3797	19	3662	*6

* $p < 0.05$ (compared to the system on its left)

Results (by error type, accuracy branch)

Error type	PBMT		Factored		NMT	
	No error	Error	No error	Error	No error	Error
Accuracy	3467	369	3525	*291	3402	266
Mistranslation	3547	289	3586	*230	3471	197
Omission	3801	35	3793	23	3619	*49
Addition	3814	22	3797	19	3655	13
Untranslated	3813	23	3797	19	3662	*6

* $p < 0.05$ (compared to the system on its left)

- Factored and NMT have less accuracy errors than PBMT
- NMT reduces untranslated (better coverage due to sub-word segmentation?)
- NMT leads to more omission errors than factored

Results (by error type, fluency branch)

Error type	PBMT		Factored		NMT	
	No error	Error	No error	Error	No error	Error
Fluency	3195	641	3298	*518	3465	**188
Unintelligible	3790	46	3769	47	3668	**0
Grammar	3270	566	3371	**445	3497	**156
Word order	3752	84	3752	64	3646	**22
Word form	3389	447	3471	*345	3538	**102
Tense...	3775	61	3765	51	3648	*20
Agreement	3466	370	3540	*276	3566	**102
Number	3778	58	3772	44	3646	*22
Gender	3788	48	3756	60	3644	*24
Case	3614	222	3694	*122	3622	**46
Person	3836	0	3816	0	3664	4

** $p < 0.01$, * $p < 0.05$ (compared to the system on its left)

Conclusions

Conclusions: contributions

1. Human fine-grained error analysis of NMT
2. NMT compared not only to pure and hierarchical PBMT but also to factored models
3. Devised an MQM-compliant taxonomy for Slavic languages
4. Approach to analyse statistically MQM results

- **Overall errors.** NMT reduces them by 54% (wrt PBMT) and by 42% (wrt factored PBMT)
- **Agreement errors** (number, gender and case). NMT is specially effective, 72% reduction (wrt PBMT), and 63% (wrt factored PBMT)
- **Omission.** The only error type for which NMT underperformed factored PBMT (40% increase)

- Compare to PBMT with morph segmentation
- NMT-focused MQM evaluation: add fine-grained tags under the Accuracy branch
- NMT vs PBMT analysis for novels

Thank you! Děkuji!

Questions?

Inter Annotator Agreement

Inter annotator agreement for each error type (min: 0.27, max: 0.72)

Error type	Cohen's κ
Accuracy	
Mistranslation	0.53
Omission	0.37
Addition	0.47
Untranslated	0.72
Fluency	
Unintelligible	0.35
Register	0.27
Word order	0.4
Function words	
Extraneous	0.46
Incorrect	0.29
Missing	0.33
Tense...	0.38
Agreement	0.33
Number	0.54
Gender	0.53
Case	0.56