The Effect of Translationese in Machine Translation Test Sets

WMT19, Florence, 2nd of August 2019

university of
groningen

Mike Zhang
Information Science Programme
University of Groningen
The Netherlands
j.j.zhang.1@student.rug.nl

Antonio Toral
CLCG
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

Overview

1. **What is translationese?**
2. **Translationese in MT data sets**
3. **Research Questions**
4. **Conclusions & Future work**

What is translationese?

Translationese

Translated text (*translationese*) $\neq$ original text

Translationese

Translated text (*translationese*) $\neq$ original text

- The differences do not indicate poor translation but rather a statistical phenomenon (Gellerstam, 1986)
- Simpler, more homogeneous, more explicit, interference from source language, aka translation universals (Baker, 1993)

Translationese in MT data sets

Translationese in MT data sets

What is the effect of translationese on MT?

- Mainly studied wrt training data (Kurokawa et al., 2009; Lembersky, 2013)

Translationese in MT data sets

What is the effect of translationese on MT?

- Mainly studied wrt training data (Kurokawa et al., 2009; Lembersky, 2013)
  - $(Source_{original}, Target_{translationese}) > (Source_{translationese}, Target_{original})$

Translationese in MT data sets

What is the effect of translationese on MT?

- Mainly studied wrt training data (Kurokawa et al., 2009; Lembersky, 2013)
    - $(Source_{original}, Target_{translationese}) > (Source_{translationese}, Target_{original})$
- Also wrt dev data, in SMT (Stymne, 2017)

Translationese in MT data sets

What is the effect of translationese on MT?

- Mainly studied wrt training data (Kurokawa et al., 2009; Lembersky, 2013)
    - $(Source_{original}, Target_{translationese}) > (Source_{translationese}, Target_{original})$
- Also wrt dev data, in SMT (Stymne, 2017)
    - Using tuning texts translated in the same original direction as the MT system tended to give a better score

Translationese in MT data sets

What is the effect of translationese on MT?

- Mainly studied wrt training data (Kurokawa et al., 2009; Lembersky, 2013)
  - $(Source_{original}, Target_{translationese}) > (Source_{translationese}, Target_{original})$
- Also wrt dev data, in SMT (Stymne, 2017)
  - Using tuning texts translated in the same original direction as the MT system tended to give a better score

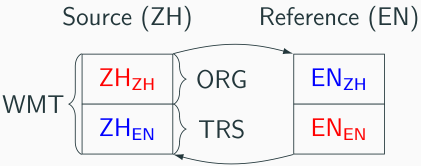
- **What about test data?**

Translationese in Test

- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)

Translationese in Test

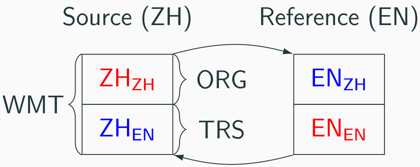
- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)



	Source (ZH)		Reference (EN)
WMT	ZH_{ZH}	ORG	EN_{ZH}
	ZH_{EN}	TRS	EN_{EN}

Translationese in Test

- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)



Source (ZH)

Reference (EN)

WMT

$\mathrm{ZH}_{\mathrm{ZH}}$

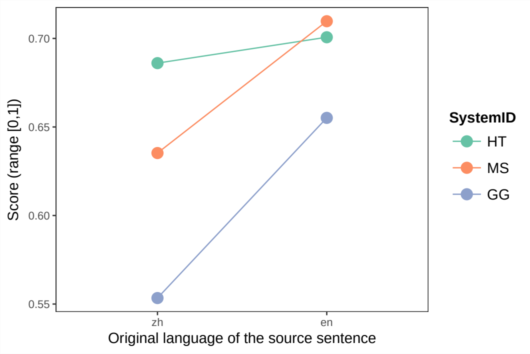
ORG

$\mathrm{EN}_{\mathrm{ZH}}$

$\mathrm{ZH}_{\mathrm{EN}}$

TRS

$\mathrm{EN}_{\mathrm{EN}}$



Translationese in Test

- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)

Translationese in Test

- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)
- Läubli et al. (2018) in similar fashion, show stronger preference for human translations over MT when evaluating documents compared to isolated sentences, on Hassan et al. (2018)

Translationese in Test

- Toral et al. (2018): translationese input favours MT systems, on Hassan et al. (2018)
- Läubli et al. (2018) in similar fashion, show stronger preference for human translations over MT when evaluating documents compared to isolated sentences, on Hassan et al. (2018)
- Taking the two works above, Graham et al. (2019) found evidence that translationese compared to original text can potentially negatively impact the accuracy of machine translation evaluations

Research Questions

Research Question(s)

1. Does the use of translationese in the source side of MT test sets unfairly favour MT systems?

Research Question(s)

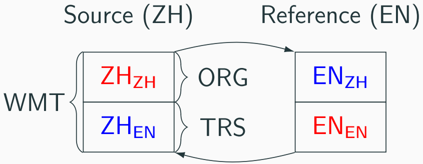
1. Does the use of translationese in the source side of MT test sets unfairly favour MT systems?
2. If the answer to RQ1 is yes, does this effect of translationese have an impact on WMT's system rankings?

Research Question(s)

1. Does the use of translationese in the source side of MT test sets unfairly favour MT systems?
2. If the answer to RQ1 is yes, does this effect of translationese have an impact on WMT's system rankings?
3. If the answer to RQ1 is yes, would some language pairs be more affected than others?

This study

- **Dataset**: WMT16, WMT17, and WMT18 → 17 translation directions, 10 unique languages (Bojar et al., 2016, 2017, 2018).
- **Human evaluation**: Direct Assessment (DA), by bilingual crowd workers and participants (Graham et al., 2013, 2014, 2017).



Source (ZH)

Reference (EN)

WMT

ZH_{ZH}

ORG

EN_{ZH}

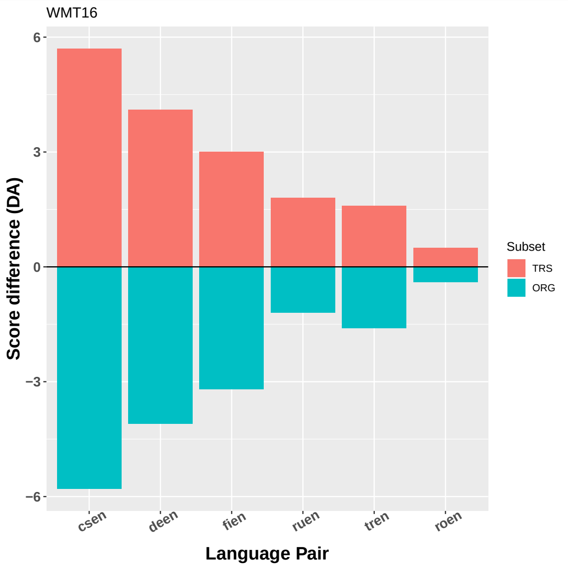
ZH_{EN}

TRS

EN_{EN}

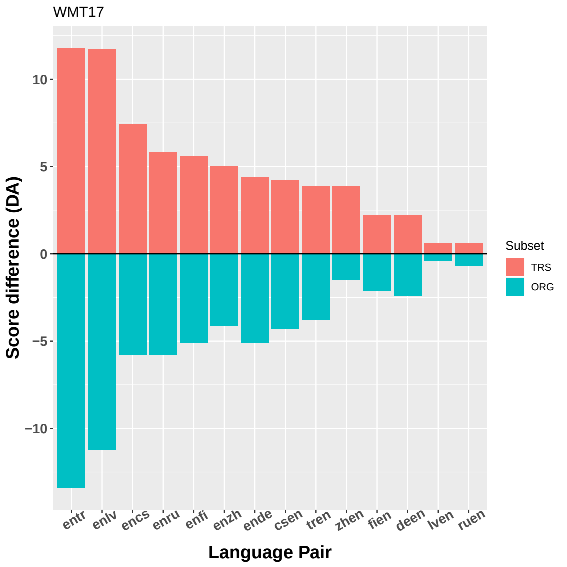
RQ1: Does Translationese Affect Human Evaluation Scores?

RQ1: favouritism for translationese, WMT16



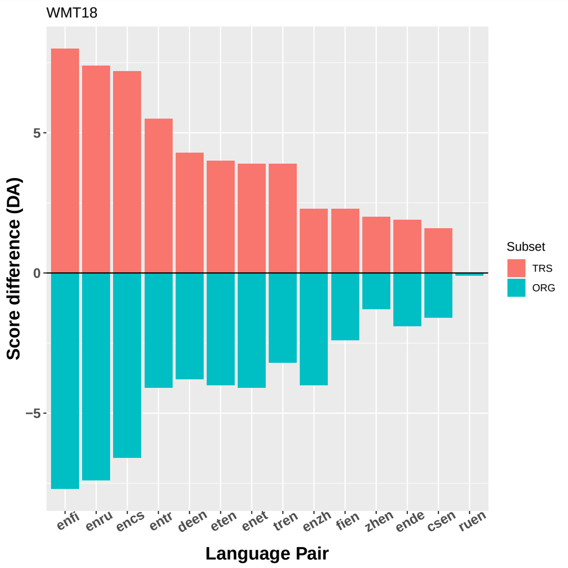
- Score difference in DA, ORG = original input, TRS = translationese input
- Consistent trend over all language pairs

WMT17



- Similar trend, TRS = inflation of scores, ORG = deflation of scores.

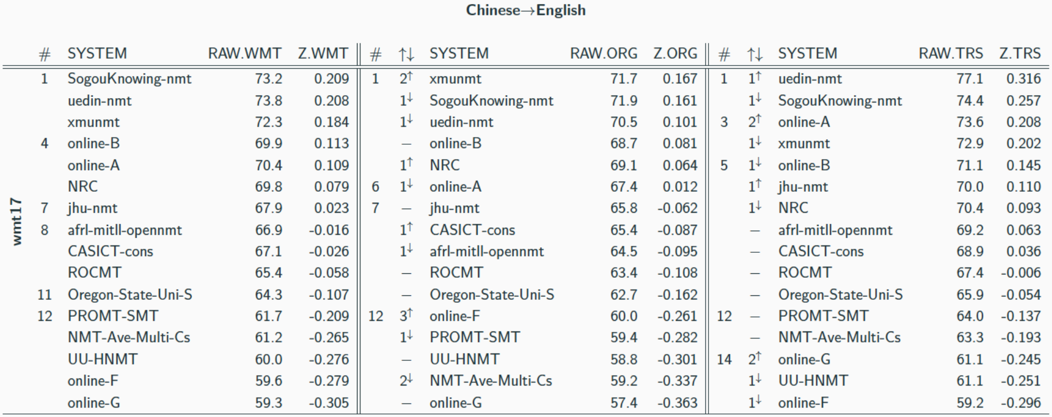
WMT18



- Again, same trend over all language pairs
- Does translationese unfairly favour MT systems?
- **Yes!**

RQ2: Do Systems' Rankings Change?

RQ2: impact on WMT’s system rankings? (e.g. ZH → EN)

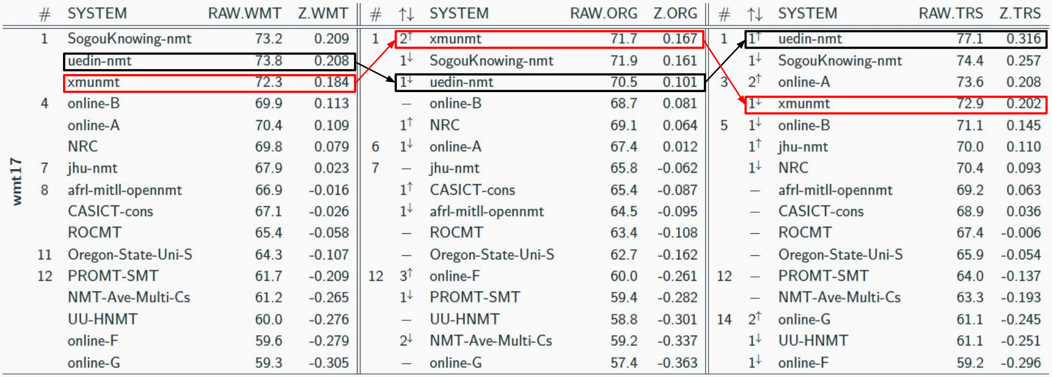


Chinese→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	SogouKnowing-nmt	73.2	0.209	1	2↑	xmunmt	71.7	0.167	1	1↑	uedin-nmt	77.1	0.316
		uedin-nmt	73.8	0.208		1↓	SogouKnowing-nmt	71.9	0.161		1↓	SogouKnowing-nmt	74.4	0.257
		xmunmt	72.3	0.184		1↓	uedin-nmt	70.5	0.101	3	2↑	online-A	73.6	0.208
	4	online-B	69.9	0.113		—	online-B	68.7	0.081		1↓	xmunmt	72.9	0.202
		online-A	70.4	0.109		1↑	NRC	69.1	0.064	5	1↓	online-B	71.1	0.145
		NRC	69.8	0.079	6	1↓	online-A	67.4	0.012		1↑	jhu-nmt	70.0	0.110
	7	jhu-nmt	67.9	0.023	7	—	jhu-nmt	65.8	-0.062		1↓	NRC	70.4	0.093
	8	afrl-mitll-opennmt	66.9	-0.016		1↑	CASICT-cons	65.4	-0.087		—	afrl-mitll-opennmt	69.2	0.063
		CASICT-cons	67.1	-0.026		1↓	afrl-mitll-opennmt	64.5	-0.095		—	CASICT-cons	68.9	0.036
		ROCMT	65.4	-0.058		—	ROCMT	63.4	-0.108		—	ROCMT	67.4	-0.006
	11	Oregon-State-Uni-S	64.3	-0.107		—	Oregon-State-Uni-S	62.7	-0.162		—	Oregon-State-Uni-S	65.9	-0.054
	12	PROMT-SMT	61.7	-0.209	12	3↑	online-F	60.0	-0.261	12	—	PROMT-SMT	64.0	-0.137
		NMT-Ave-Multi-Cs	61.2	-0.265		1↓	PROMT-SMT	59.4	-0.282		—	NMT-Ave-Multi-Cs	63.3	-0.193
		UU-HNMT	60.0	-0.276		—	UU-HNMT	58.8	-0.301	14	2↑	online-G	61.1	-0.245
		online-F	59.6	-0.279		2↓	NMT-Ave-Multi-Cs	59.2	-0.337		1↓	UU-HNMT	61.1	-0.251
		online-G	59.3	-0.305		—	online-G	57.4	-0.363		1↓	online-F	59.2	-0.296

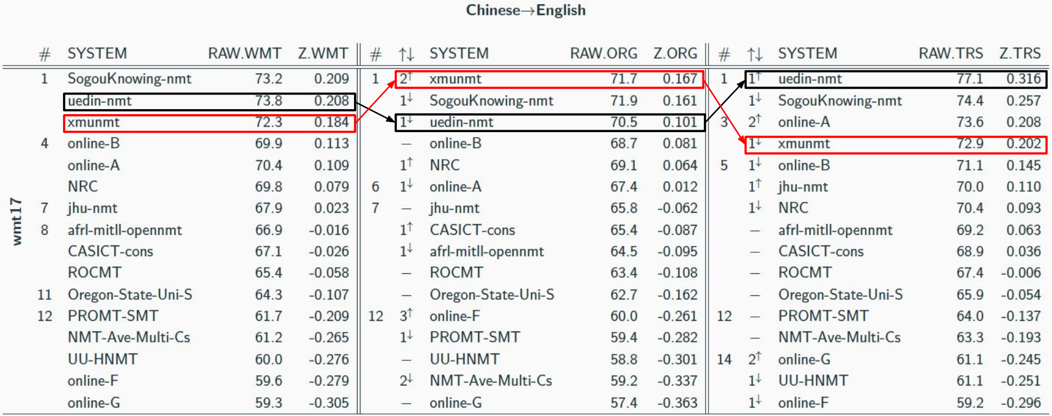
RQ2: impact on WMT's system rankings? (e.g. ZH → EN)

Chinese→English



	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	SogouKnowing-nmt	73.2	0.209	1	2↑	xmunmt	71.7	0.167	1	1↑	uedin-nmt	77.1	0.316
		uedin-nmt	73.8	0.208		1↓	SogouKnowing-nmt	71.9	0.161		1↓	SogouKnowing-nmt	74.4	0.257
		xmunmt	72.3	0.184		1↓	uedin-nmt	70.5	0.101	3	2↑	online-A	73.6	0.208
	4	online-B	69.9	0.113		—	online-B	68.7	0.081		1↓	xmunmt	72.9	0.202
		online-A	70.4	0.109		1↑	NRC	69.1	0.064	5	1↓	online-B	71.1	0.145
		NRC	69.8	0.079	6	1↓	online-A	67.4	0.012		1↑	jhu-nmt	70.0	0.110
	7	jhu-nmt	67.9	0.023	7	—	jhu-nmt	65.8	-0.062		1↓	NRC	70.4	0.093
	8	afrl-mitll-opennmt	66.9	-0.016		1↑	CASICT-cons	65.4	-0.087		—	afrl-mitll-opennmt	69.2	0.063
		CASICT-cons	67.1	-0.026		1↓	afrl-mitll-opennmt	64.5	-0.095		—	CASICT-cons	68.9	0.036
		ROCMT	65.4	-0.058		—	ROCMT	63.4	-0.108		—	ROCMT	67.4	-0.006
	11	Oregon-State-Uni-S	64.3	-0.107		—	Oregon-State-Uni-S	62.7	-0.162		—	Oregon-State-Uni-S	65.9	-0.054
	12	PROMT-SMT	61.7	-0.209	12	3↑	online-F	60.0	-0.261	12	—	PROMT-SMT	64.0	-0.137
		NMT-Ave-Multi-Cs	61.2	-0.265		1↓	PROMT-SMT	59.4	-0.282		—	NMT-Ave-Multi-Cs	63.3	-0.193
		UU-HNMT	60.0	-0.276		—	UU-HNMT	58.8	-0.301	14	2↑	online-G	61.1	-0.245
		online-F	59.6	-0.279		2↓	NMT-Ave-Multi-Cs	59.2	-0.337		1↓	UU-HNMT	61.1	-0.251
		online-G	59.3	-0.305		—	online-G	57.4	-0.363		1↓	online-F	59.2	-0.296

RQ2: impact on WMT's system rankings? (e.g. ZH → EN)

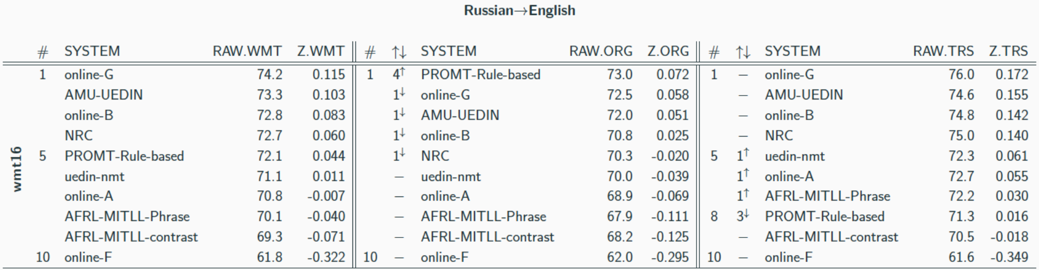


Chinese→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	SogouKnowing-nmt	73.2	0.209	1	2↑	xmunmt	71.7	0.167	1	1↑	uedin-nmt	77.1	0.316
		uedin-nmt	73.8	0.208		1↓	SogouKnowing-nmt	71.9	0.161		1↓	SogouKnowing-nmt	74.4	0.257
		xmunmt	72.3	0.184		1↓	uedin-nmt	70.5	0.101	3	2↑	online-A	73.6	0.208
	4	online-B	69.9	0.113		—	online-B	68.7	0.081		1↓	xmunmt	72.9	0.202
		online-A	70.4	0.109		1↑	NRC	69.1	0.064	5	1↓	online-B	71.1	0.145
		NRC	69.8	0.079	6	1↓	online-A	67.4	0.012		1↑	jhu-nmt	70.0	0.110
	7	jhu-nmt	67.9	0.023	7	—	jhu-nmt	65.8	-0.062		1↓	NRC	70.4	0.093
	8	afrl-mitll-opennmt	66.9	-0.016		1↑	CASICT-cons	65.4	-0.087		—	afrl-mitll-opennmt	69.2	0.063
		CASICT-cons	67.1	-0.026		1↓	afrl-mitll-opennmt	64.5	-0.095		—	CASICT-cons	68.9	0.036
		ROCMT	65.4	-0.058		—	ROCMT	63.4	-0.108		—	ROCMT	67.4	-0.006
	11	Oregon-State-Uni-S	64.3	-0.107		—	Oregon-State-Uni-S	62.7	-0.162		—	Oregon-State-Uni-S	65.9	-0.054
	12	PROMT-SMT	61.7	-0.209	12	3↑	online-F	60.0	-0.261	12	—	PROMT-SMT	64.0	-0.137
		NMT-Ave-Multi-Cs	61.2	-0.265		1↓	PROMT-SMT	59.4	-0.282		—	NMT-Ave-Multi-Cs	63.3	-0.193
		UU-HNMT	60.0	-0.276		—	UU-HNMT	58.8	-0.301	14	2↑	online-G	61.1	-0.245
		online-F	59.6	-0.279		2↓	NMT-Ave-Multi-Cs	59.2	-0.337		1↓	UU-HNMT	61.1	-0.251
		online-G	59.3	-0.305		—	online-G	57.4	-0.363		1↓	online-F	59.2	-0.296

- Clusters change: WMT(1,4,7,8,11,12)→ORG(1,6,7,12)→TRS(1,3,5,12,14)

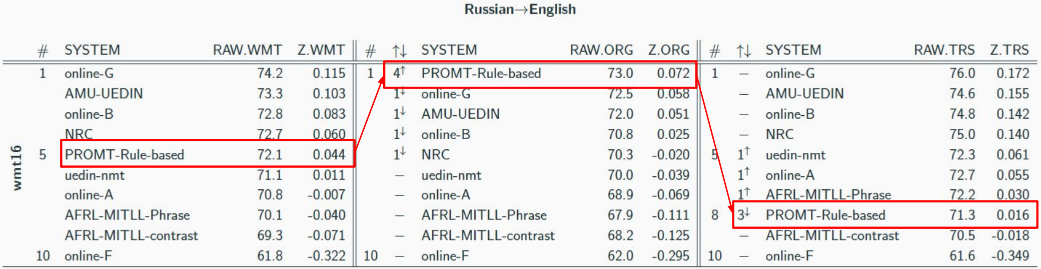
Another example (RU → EN)



Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	4↑	PROMT-Rule-based	73.0	0.072	1	–	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		1↓	online-G	72.5	0.058		–	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		1↓	AMU-UEDIN	72.0	0.051		–	online-B	74.8	0.142
		NRC	72.7	0.060		1↓	online-B	70.8	0.025		–	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		1↓	NRC	70.3	-0.020	5	1↑	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		–	uedin-nmt	70.0	-0.039		1↑	online-A	72.7	0.055
		online-A	70.8	-0.007		–	online-A	68.9	-0.069		1↑	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		–	AFRL-MITLL-Phrase	67.9	-0.111	8	3↓	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		–	AFRL-MITLL-contrast	68.2	-0.125		–	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	–	online-F	62.0	-0.295	10	–	online-F	61.6	-0.349

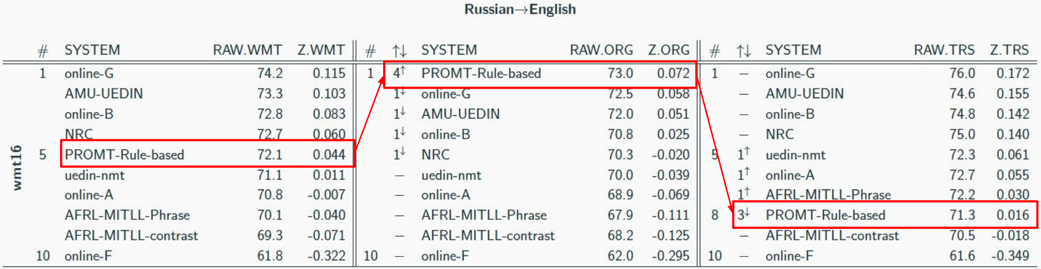
Another example (RU → EN)



Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	$4^{\uparrow}$	PROMT-Rule-based	73.0	0.072	1	—	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		$1^{\downarrow}$	online-G	72.5	0.058		—	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		$1^{\downarrow}$	AMU-UEDIN	72.0	0.051		—	online-B	74.8	0.142
		NRC	72.7	0.060		$1^{\downarrow}$	online-B	70.8	0.025		—	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		$1^{\downarrow}$	NRC	70.3	-0.020	5	$1^{\uparrow}$	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		—	uedin-nmt	70.0	-0.039		$1^{\uparrow}$	online-A	72.7	0.055
		online-A	70.8	-0.007		—	online-A	68.9	-0.069		$1^{\uparrow}$	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		—	AFRL-MITLL-Phrase	67.9	-0.111	8	$3^{\downarrow}$	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		—	AFRL-MITLL-contrast	68.2	-0.125		—	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	—	online-F	62.0	-0.295	10	—	online-F	61.6	-0.349

Another example (RU → EN)

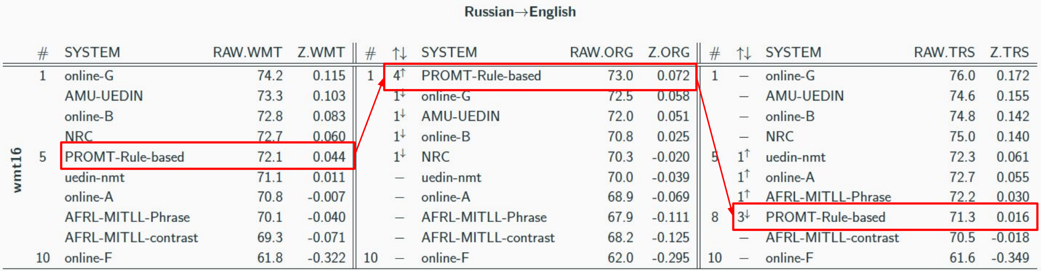


Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	$4^{\uparrow}$	PROMT-Rule-based	73.0	0.072	1	—	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		$1^{\downarrow}$	online-G	72.5	0.058		—	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		$1^{\downarrow}$	AMU-UEDIN	72.0	0.051		—	online-B	74.8	0.142
		NRC	72.7	0.060		$1^{\downarrow}$	online-B	70.8	0.025		—	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		$1^{\downarrow}$	NRC	70.3	-0.020	5	$1^{\uparrow}$	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		—	uedin-nmt	70.0	-0.039		$1^{\uparrow}$	online-A	72.7	0.055
		online-A	70.8	-0.007		—	online-A	68.9	-0.069		$1^{\uparrow}$	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		—	AFRL-MITLL-Phrase	67.9	-0.111	8	$3^{\downarrow}$	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		—	AFRL-MITLL-contrast	68.2	-0.125		—	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	—	online-F	62.0	-0.295	10	—	online-F	61.6	-0.349

- Clusters change: WMT(1,5,10)→ORG(1,10)→TRS(1,5,8,10)

Another example (RU → EN)



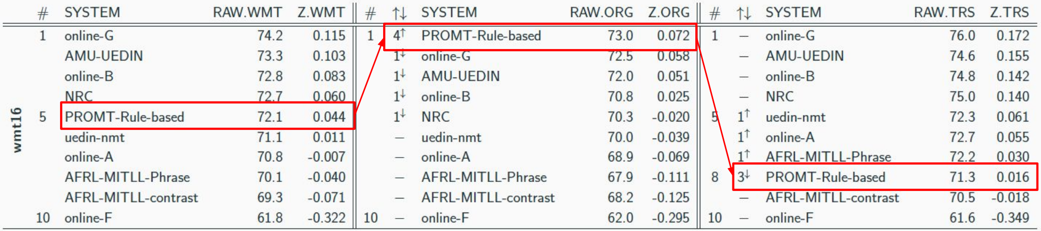
Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	$4^{\uparrow}$	PROMT-Rule-based	73.0	0.072	1	–	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		$1^{\downarrow}$	online-G	72.5	0.058		–	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		$1^{\downarrow}$	AMU-UEDIN	72.0	0.051		–	online-B	74.8	0.142
		NRC	72.7	0.060		$1^{\downarrow}$	online-B	70.8	0.025		–	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		$1^{\downarrow}$	NRC	70.3	-0.020	5	$1^{\uparrow}$	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		–	uedin-nmt	70.0	-0.039		$1^{\uparrow}$	online-A	72.7	0.055
		online-A	70.8	-0.007		–	online-A	68.9	-0.069		$1^{\uparrow}$	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		–	AFRL-MITLL-Phrase	67.9	-0.111	8	$3^{\downarrow}$	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		–	AFRL-MITLL-contrast	68.2	-0.125		–	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	–	online-F	62.0	-0.295	10	–	online-F	61.6	-0.349

- Clusters change: WMT(1,5,10)→ORG(1,10)→TRS(1,5,8,10)
- So would there be ranking changes?

Another example (RU → EN)

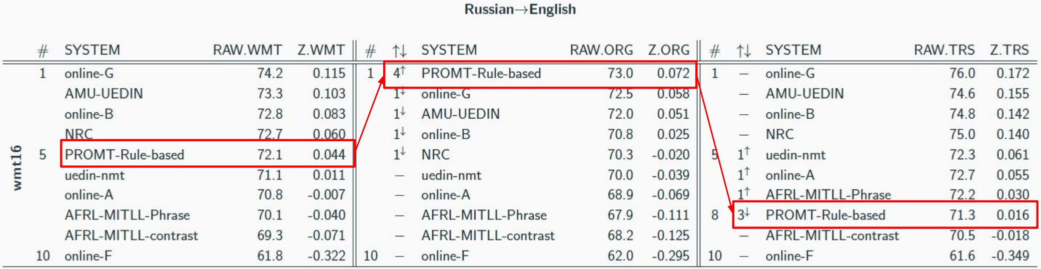
Russian→English



	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	4↑	PROMT-Rule-based	73.0	0.072	1	–	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		1↓	online-G	72.5	0.058		–	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		1↓	AMU-UEDIN	72.0	0.051		–	online-B	74.8	0.142
		NRC	72.7	0.060		1↓	online-B	70.8	0.025		–	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		1↓	NRC	70.3	-0.020	5	1↑	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		–	uedin-nmt	70.0	-0.039		1↑	online-A	72.7	0.055
		online-A	70.8	-0.007		–	online-A	68.9	-0.069		1↑	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		–	AFRL-MITLL-Phrase	67.9	-0.111	8	3↓	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		–	AFRL-MITLL-contrast	68.2	-0.125		–	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	–	online-F	62.0	-0.295	10	–	online-F	61.6	-0.349

- Clusters change: WMT(1,5,10)→ORG(1,10)→TRS(1,5,8,10)
- So would there be ranking changes?
- **Yes, and clusters too!**

Another example (RU → EN)



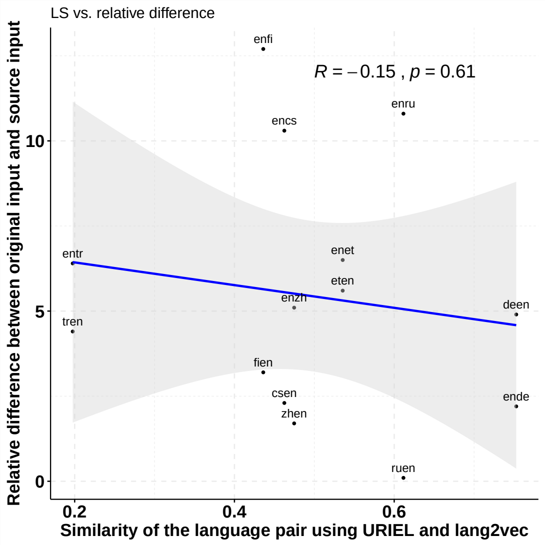
Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	$4^{\uparrow}$	PROMT-Rule-based	73.0	0.072	1	–	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		$1^{\downarrow}$	online-G	72.5	0.058		–	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		$1^{\downarrow}$	AMU-UEDIN	72.0	0.051		–	online-B	74.8	0.142
		NRC	72.7	0.060		$1^{\downarrow}$	online-B	70.8	0.025		–	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		$1^{\downarrow}$	NRC	70.3	-0.020	5	$1^{\uparrow}$	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		–	uedin-nmt	70.0	-0.039		$1^{\uparrow}$	online-A	72.7	0.055
		online-A	70.8	-0.007		–	online-A	68.9	-0.069		$1^{\uparrow}$	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		–	AFRL-MITLL-Phrase	67.9	-0.111	8	$3^{\downarrow}$	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		–	AFRL-MITLL-contrast	68.2	-0.125		–	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	–	online-F	62.0	-0.295	10	–	online-F	61.6	-0.349

- Clusters change: WMT(1,5,10)→ORG(1,10)→TRS(1,5,8,10)
- So would there be ranking changes?
- **Yes, and clusters too!**
- However, half data

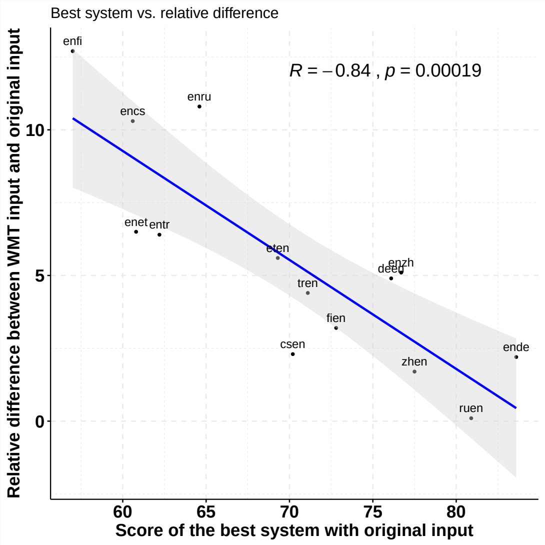
RQ3: Are Some Languages More Affected?

Research Question 3: is there a trend?



- Language similarity (lang2vec (Littell et al., 2017)) vs. relative difference between WMT input and ORG input
- Low correlation

Research Question 3: is there a trend?



- Highest scoring system (with only ORG input) vs. relative difference between WMT input and ORG input
- High correlation!
- High differences could be due to under-resourced languages

Conclusions & Future work

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.
- **Translation quality**:

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.
- **Translation quality**:
  - Correlation between the effect of translationese and the translation quality attainable for translation directions.

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.
- **Translation quality**:
  - Correlation between the effect of translationese and the translation quality attainable for translation directions.
  - The effect of translationese tends to be high when an under-resourced language is present.

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.
- **Translation quality**:
  - Correlation between the effect of translationese and the translation quality attainable for translation directions.
  - The effect of translationese tends to be high when an under-resourced language is present.
- **Recommendations (?)**: the WMT organizers have addressed this issue by providing completely source-language native test sets for WMT19.

Conclusion

- **Translationese**: if present, it inflates DA scores. If removed, it lowers DA scores.
- **Translation quality**:
    - Correlation between the effect of translationese and the translation quality attainable for translation directions.
    - The effect of translationese tends to be high when an under-resourced language is present.
- **Recommendations (?)**: the WMT organizers have addressed this issue by providing completely source-language native test sets for WMT19.
- **Future work**: characteristics of translationese in the WMT test sets.

Ack. WMT: for providing the data

Ack. WMT: for providing the data

Thank you!

Questions?

Mike Zhang & Antonio Toral

j.j.zhang.1@student.rug.nl — a.toral.ruiz@rug.nl

References

M. Baker. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250, 1993.

O. Bojar et al. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198, 2016.

References ii

O. Bojar et al. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, 2017. URL `http://www.statmt.org/wmt17/pdf/WMT17.pdf`.

O. Bojar et al. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–303, 2018. URL `http://aclweb.org/anthology/W18-6401.pdf`.

M. Gellerstam. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95, 1986.

Y. Graham, B. Haddow, and P. Koehn. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*, 2019.

References iii

Y. Graham et al. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, 2013.

Y. Graham et al. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, 2014.

Y. Graham et al. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30, 2017.

References iv

H. Hassan et al. Achieving Human Parity on Automatic Chinese to English News Translation. 2018. URL `https://www.microsoft.com/en-us/research/publication/achieving-human-parity-on-automatic-chinese-to-english-news-transla` `https://arxiv.org/abs/1803.05567`.

D. Kurokawa et al. Automatic detection of translated text and its impact on machine translation. *Proceedings of MT-Summit XII*, pages 81–88, 2009. URL `https://arxiv.org/pdf/1808.07048.pdf`.

S. Läubli, R. Sennrich, and M. Volk. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*, 2018. URL `https://arxiv.org/pdf/1808.07048.pdf`.

References v

G. Lembersky. *The Effect of Translationese on Statistical Machine Translation*. University of Haifa, Faculty of Social Sciences, Department of Computer Science, 2013.

P. Littell et al. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, 2017.

S. Stymne. The effect of translationese on tuning for statistical machine translation. In *The 21st Nordic Conference on Computational Linguistics*, pages 241–246, 2017.

References vi

A. Toral et al. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*, 2018. URL `https://arxiv.org/abs/1808.10432`.

	With Ties			Mean		Without Ties			
Language Direction	WMT16	WMT17	WMT18			WMT16	WMT17	WMT18	Language Direction
Romanian → English†	1.000*	-	-	1.000	1.000	1.000*	-	-	Romanian → English †
Turkish → English	0.983*	0.948*	1.000*	0.977	1.000	1.000*	1.000*	1.000*	Czech → English
Finnish → English	0.943*	0.966*	1.000*	0.970	0.978	-	-	0.978*	English → Estonian †
Czech → English	0.929*	1.000*	0.949*	0.959	0.956	-	-	0.956*	Estonian → English †
German → English	0.979*	0.939*	0.906*	0.941	0.944	-	0.944*	-	Latvian → English †
English → Czech	-	0.904*	0.949*	0.927	0.929	-	0.929*	0.929*	English → Turkish
Latvian → English†	-	0.921*	-	0.921	0.917	-	0.889*	0.944*	English → Russian
English → Finnish	-	0.868*	0.968*	0.918	0.898	-	0.927*	0.868*	English → Chinese
English → Russian	-	0.873*	0.935*	0.904	0.882	-	0.882*	-	English → Latvian †
Chinese → English	-	0.923*	0.882*	0.903	0.869	0.733*	0.944*	0.929*	Russian → English
English → German	-	0.863*	0.856*	0.860	0.852	1.000*	1.000*	0.556*	Finnish → English
English → Estonian†	-	-	0.845*	0.845	0.848	0.833*	0.911*	0.800*	Turkish → English
Estonian → English†	-	-	0.830*	0.830	0.784	-	0.633*	0.934*	Chinese → English
English → Chinese	-	0.847*	0.789*	0.818	0.726	-	0.451*	1.000*	English → Czech
English → Turkish	-	0.890*	0.734*	0.812	0.713	0.911*	0.345	0.883*	German → English
Russian → English	0.557	0.845*	0.890*	0.764	0.675	-	0.817*	0.533*	English → German
English → Latvian †	-	0.718*	-	0.718	0.637	-	0.970*	0.303	English → Finnish