Proposals for master theses

Academic year 2018-19

Antonio Toral
https://antoniotor.al
a.toral.ruiz@rug.nl

University of Groningen

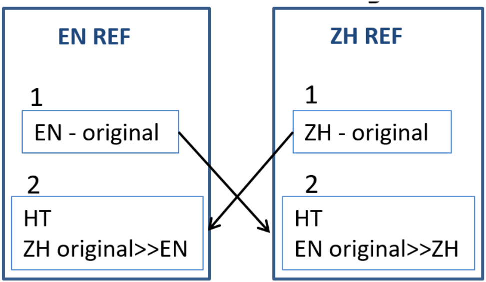
November 9, 2018

Proposals

1. Translationese in MT Testsets
2. Posteditese
3. MT of Noisy Input

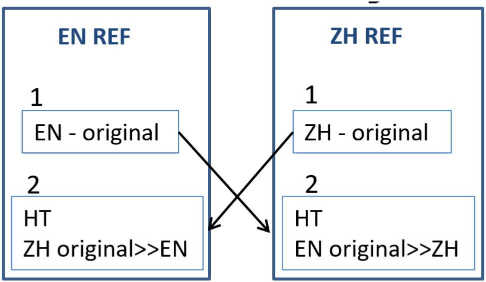
Problem

Test sets at WMT are symmetrical



Problem

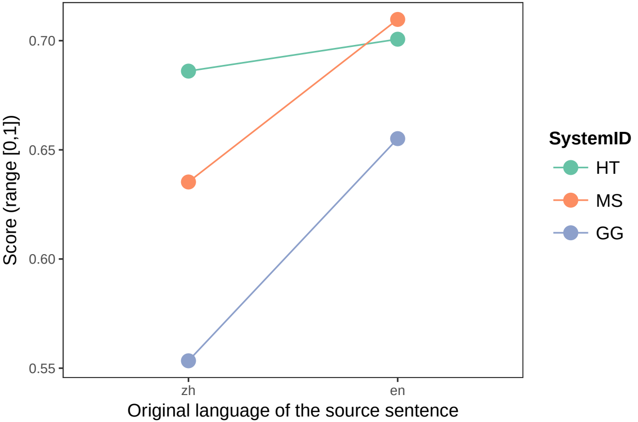
Test sets at WMT are symmetrical



Why a problem: the translationese part may be artificially easier for MT due to 3 principles of translationese: simplification, explicitation and normalisation.

Problem

EN–ZH WMT2017 testset



RQs

- RQ1. Is this issue present in other datasets or is it just an artifact of the EN–ZH 2017?
- RQ2. Would removing translationese change the system rankings?

RQs

- RQ1. Is this issue present in other datasets or is it just an artifact of the EN–ZH 2017?
- RQ2. Would removing translationese change the system rankings?

Czech→English

	Ave. %	Ave. z	System
1	71.8	0.298	CUNI-TRANSFORMER
2	67.9	0.165	UEDIN
3	66.6	0.115	ONLINE-B
4	62.1	−0.023	ONLINE-A
5	57.5	−0.183	ONLINE-G

English→Czech

	Ave. %	Ave. z	System
1	67.2	0.594	CUNI-TRANSFORMER
2	60.6	0.384	UEDIN
3	52.1	0.101	ONLINE-B
4	46.0	−0.115	ONLINE-A
5	42.0	−0.246	ONLINE-G

RQs

- RQ1. Is this issue present in other datasets or is it just an artifact of the EN–ZH 2017?
- RQ2. Would removing translationese change the system rankings?
- RQ3. Are some language pairs (e.g. more related) or some systems (e.g. SMT) more affected than others?
- RQ4. What are the characteristics of translationese? Syntax, vocabulary variety, etc.

What is there

- Testsets for WMT 2006 to 2018, at least 3 language pairs per year
- MT outputs for the testsets
- Scripts for analysing translationese in EN–ZH 2017 [Toral et al., 2018]

Plan

Translationese in MT Testsets

Posteditese

MT of Noisy Input

Problem

3 types of translations:

1. Human (from scratch)
2. Machine (automatic)
3. Machine-assisted (post-edited)

Problem

3 types of translations:

1. Human (from scratch)
2. Machine (automatic)
3. Machine-assisted (post-edited)

Can they be distinguished from each other? Or: can we build an effective binary classifier?

- Yes, between 1 and 2
- Not yet, between 1 and 3

Problem: from scratch vs posteditese

In theory, posteditese should be distinguisable from translations from scratch

In practice: [Daems et al., 2017] achieved 50% accuracy

- Very small dataset: 8 articles, 160 words each (EN–NL)

Other available datasets:

- TED talks EN–FR and EN–DE. 600 sentences each
- Novel EN–CA. 330 sentences
- Industry data?

Plan

Translationese in MT Testsets

Posteditese

MT of Noisy Input

Problem

Jointly with Rob and Gertjan

[Michel and Neubig, 2018] introduces a corpus of noisy input and translations thereof. EN–{FR, JA}.

- Train: 6K to 36K sentences
- Test: 1K

Their approach

- Train on clean data
- Fine-tune on noisy input

Issue: vocabulary mismatch

Idea

Use MoNoise [van der Goot and van Noord, 2017]

- Clean the noisy data with MoNoise
- Train MT using clean and cleaned data

More advanced possibilities

- Give the n-best output of MoNoise to NMT [van der Goot and van Noord, 2018]
- Learn jointly MoNoise and NMT

References I

Daems, J., Clercq, O. D., and Macken, L. (2017).
Translationese and Post-editese: How comparable is comparable quality?

Michel, P. and Neubig, G. (2018).
Mtnt: A testbed for machine translation of noisy text.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553. Association for Computational Linguistics.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018).
Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation.
In *Proceedings of WMT*, pages 113–123, Brussels, Belgium.

van der Goot, R. and van Noord, G. (2017).
Monoise: Modeling noise using a modular normalization system.
CoRR, abs/1710.03476.

References II

van der Goot, R. and van Noord, G. (2018).
Modeling input uncertainty in neural network dependency parsing.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4991. Association for Computational Linguistics.

Thank you!

Questions?