APE at Scale and its Implications on MT Evaluation Biases

Markus Freitag, Isaac Caswell, Scott Roy

Google Research

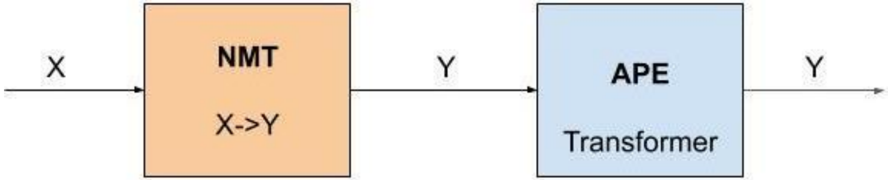
Talk Introduction

Automatic Post-Editing at Scale

Generate synthetic APE training data

Table-to-text SportResults task

WMT news translation task



MT Evaluation Biases

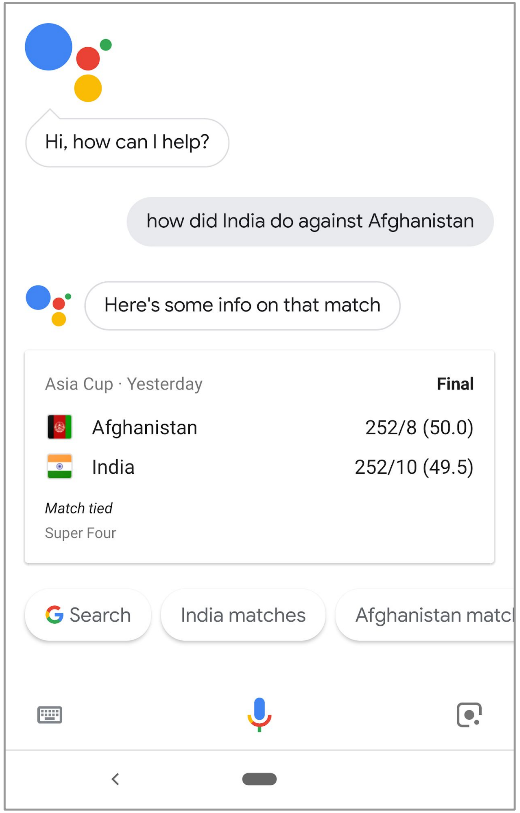
Use APE output as a tool to investigate the effect of translationese on MT evaluation

John Dryden wrote about translationese as early as 1685.

Motivation

or why we did start training APE models
or why we treat NMT as a black box

Motivation - NLG

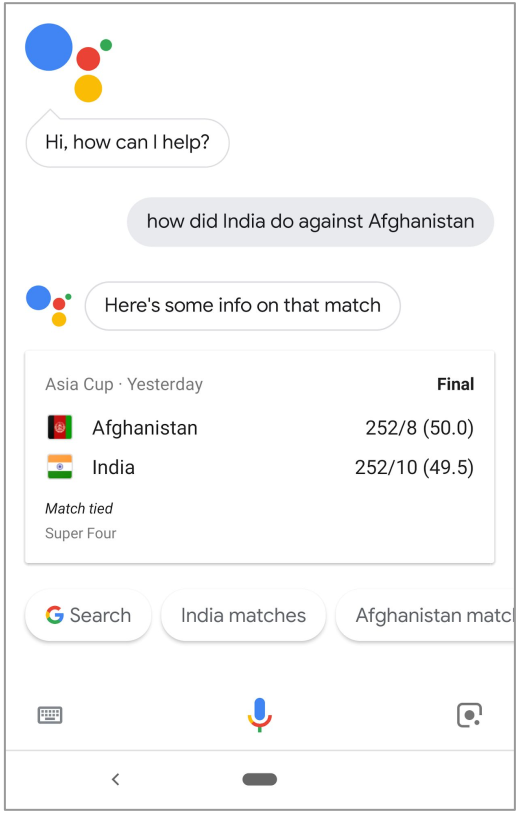


India versus Afghanistan ended in a tie. India scored 252 all-out in 49.5 overs and Afghanistan scored 252 for 8 in 50 overs.

Google Translate

Indien gegen Afghanistan endete unentschieden. Indien erzielte 252 All-Out in 49,5 Overs und Afghanistan 252 für 8 in 50 Overs.

Motivation - NLG



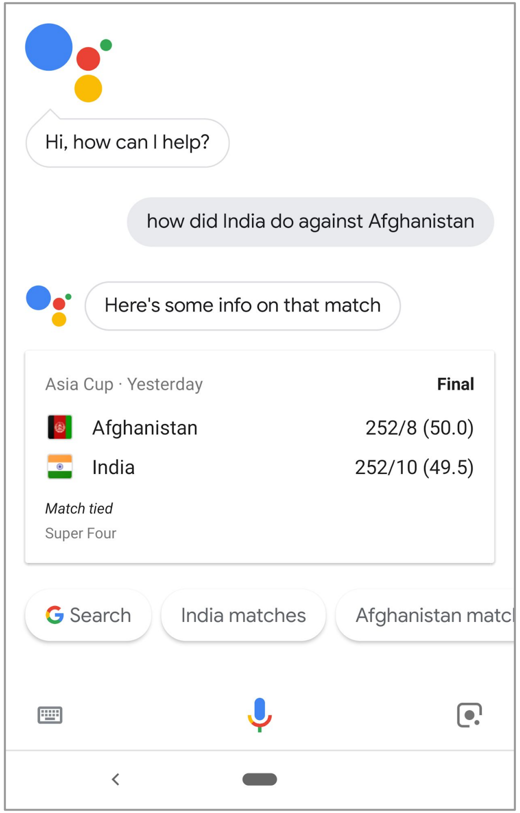
India versus Afghanistan ended in a tie. India scored 252 all-out in 49.5 overs and Afghanistan scored 252 for 8 in 50 overs.

Google Translate

accurate

Indien gegen Afghanistan endete unentschieden. Indien erzielte 252 All-Out in 49,5 Overs und Afghanistan 252 für 8 in 50 Overs.

Motivation - NLG



India versus Afghanistan ended in a tie. India scored 252 all-out in 49.5 overs and Afghanistan scored 252 for 8 in 50 overs.

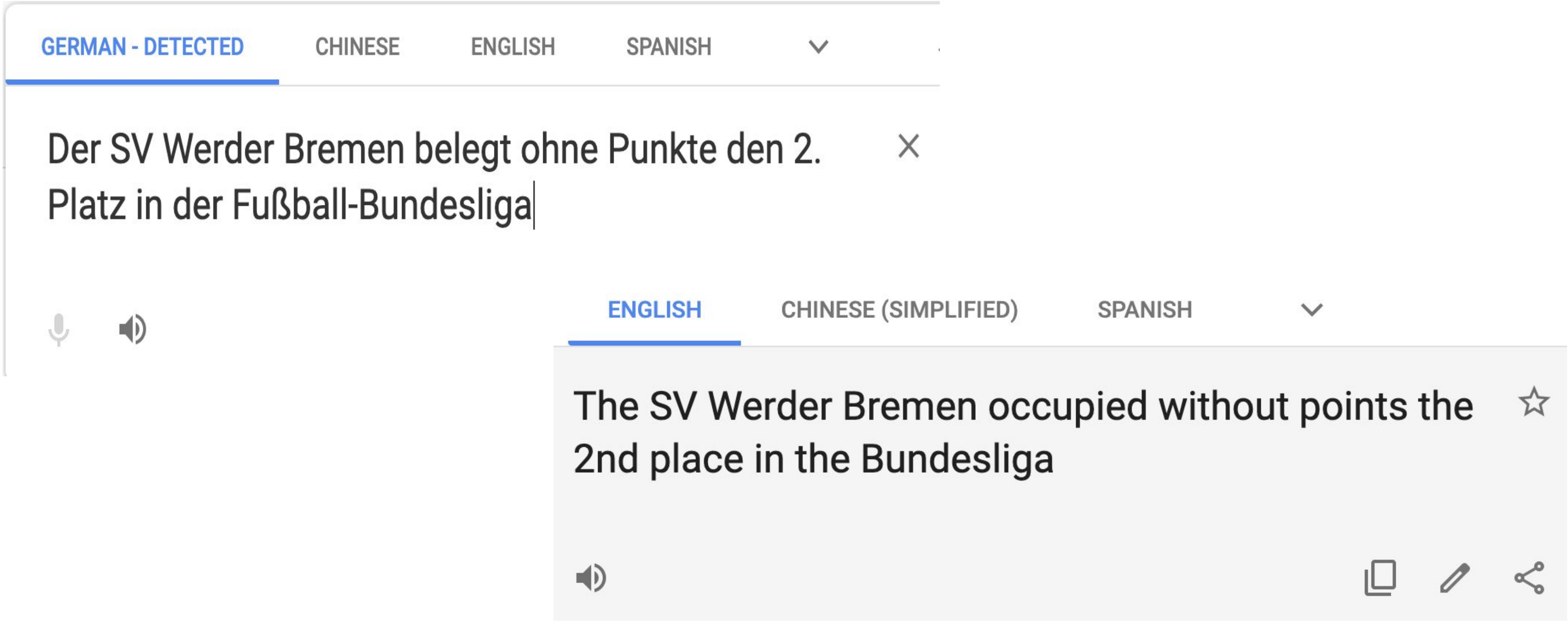
Google Translate

accurate

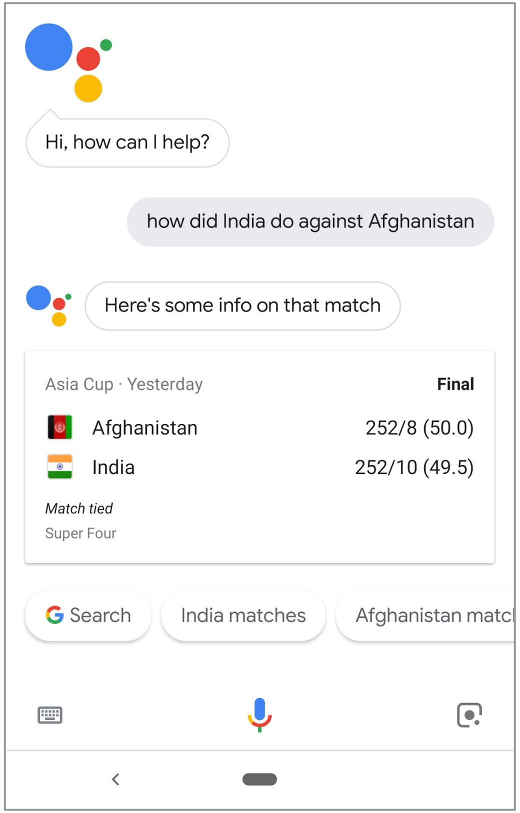
Not fluent

Indien gegen Afghanistan endete unentschieden. Indien erzielte 252 All-Out in 49,5 Overs und Afghanistan 252 für 8 in 50 Overs.

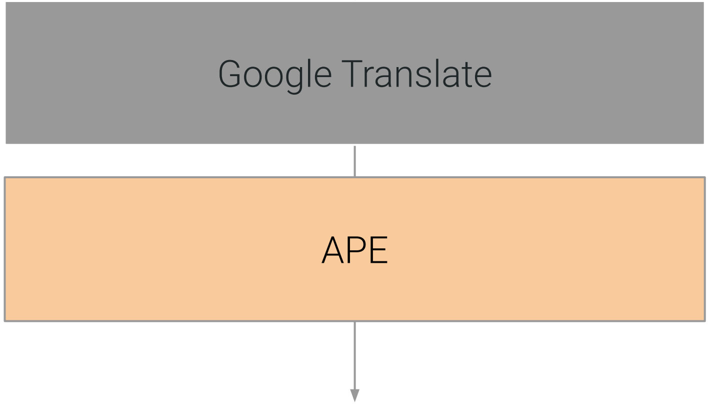
Example - Why we need APE



Motivation - NLG



India versus Afghanistan ended in a tie. India scored 252 all-out in 49.5 overs and Afghanistan scored 252 for 8 in 50 overs.

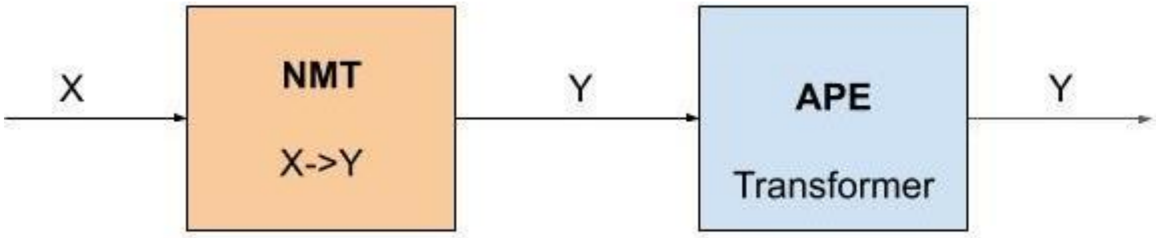


Indien spielte unentschieden gegen Afghanistan. Indien erzielte 252 All-outs in 49,5 Overs. Afghanistan erzielte 252 All-Outs für 8 Wickets in 50 Overs.

Automatic Post Editing

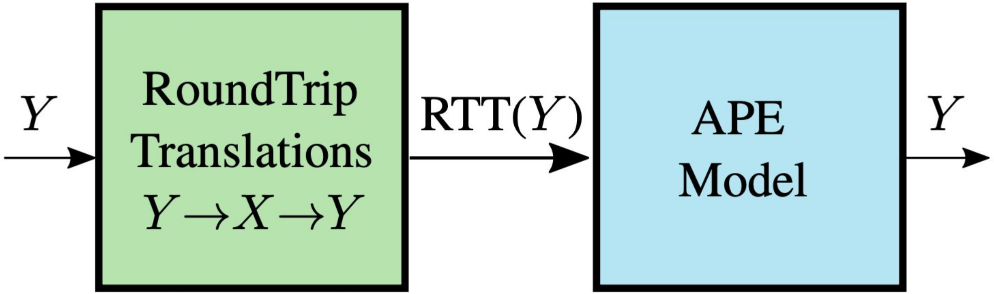
- Idea of Automatic Post Editing (APE):
    - Automatically improve noisy MT output into natural and accurate text
    - Train a second model that “translates” a noisy sentence into a clean sentence

- APE without human post-edited data:
    - Use only unlabeled high quality data
    - Model noisy MT output with roundtrip translations (RTT) (Junczys-Dowmunt et al)
    - Train a transformer model on pairs of RTT(M)->M

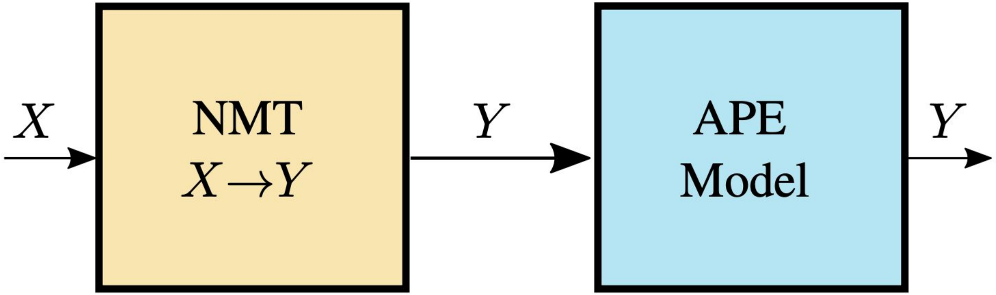


APE trained on Synthetic Data

Training:



Inference:

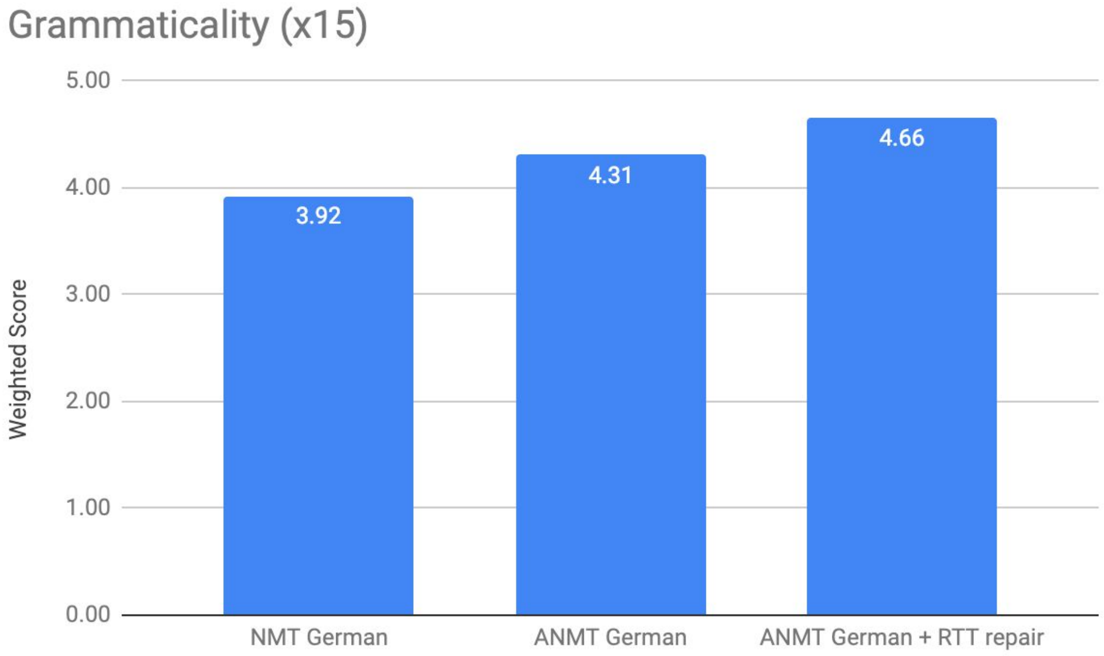


Experimental Setup

- Domain: Sport Results
- Underlying NMT model: Google Translate adapted to the assistant domain
- APE model trained on 5M in-domain sentences
  - Sentences are crawled from the web and filtered by
    - Language-id
    - Entity filtering (all target sentences have at least one sport team)
    - Ranked by a language model trained on 2k human translated data

In-domain data not part of the NMT model

NLG - Human Evaluation



Example

<u>Source:</u>
The Houston Rockets are leading the Western Conference.

<u>Google Translate:</u>
Die Houston Rockets sind die führende westliche Konferenz.

<u>Google Translate + APE:</u>
Die Houston Rockets führen die Western Conference an.

Example - much nicer!

<u>Source:</u>
SV Werder Bremen is second in the German football Bundesliga without a point.

<u>Google Translate:</u>
Der SV Werder Bremen **steht** ohne Punkte **auf dem** 2. Platz der Fußball Bundesliga.

<u>Google Translate + APE:</u>
Der SV Werder Bremen **belegt** ohne Punkte **den** 2. Platz in der Fußball-Bundesliga.

APE in the WMT News Setting

MT is not a black box anymore

APE for General NMT

- Can we also improve general NMT with APE?
- Sport Results was a very tight domain
- Experimental setup:
    - WMT news domain English->{German, Romanian, French}
        - Still not open domain
        - But much larger
    - NMT systems either only trained on bitext or with noised-back-translation (NBT)

!! MT is no longer a black box !!

Baseline NMT Systems

- NMT systems either trained on
  - **Bitext**
    - No overlap between APE training data and NMT training data
    - Question: Is APE an alternative to BT?
  - **Noised Back-Translation**
    - Same monolingual corpus for both NBT and APE
    - Question: APE helpful on top of BT?

Background: (Noised) Back-Translation

- **Back-Translation (BT)** has proven one of the simplest and most effective ways to use monolingual data for NMT
- **Noised Back-Translation** *(Edunov et al. 2018, Imamura et al. 2018)* has shown large gains over standard BT on WMT EnDe and EnFr
- Noise can be from sampled decoding or heuristic noise as in Lample et al.
    - Lample noise: Word-dropout, word-blanking, constrained permutation

Noise type	Example sentence
[no noise]	Raise the child, love the child.
NoisedBT	Raise child ___ love child, the.

Training Data Statistics

3 different scenarios:

- Intermediate (bitext) - large (monolingual)
- Large (bitext) - large (monolingual)
- Tiny (bitext) - small (monolingual)

	bitext	monolingual
WMT18 English->German	5M	216.5M
WMT14 English->French	41M	34M
WMT16 English->Romanian	0.5M	2.2M

Results are Underwhelming?

	newstest2014	newstest2015	newstest2016	newstest2017	average
Vaswani et al. (2017)	28.4	-	-	-	
Shaw et al. (2018)	29.2	-	-	-	
our bitext	29.2	31.4	35.0	29.4	31.2
our NBT	33.5	34.4	38.3	32.5	34.7

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Results are Underwhelming?

	newstest2014	newstest2015	newstest2016	newstest2017	average
Vaswani et al. (2017)	28.4	-	-	-	
Shaw et al. (2018)	29.2	-	-	-	
our bitext	29.2	31.4	35.0	29.4	31.2
+ RTT APE	30.7	31.2	33.6	30.1	31.4
our NBT	33.5	34.4	38.3	32.5	34.7

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Results are Underwhelming?

	newstest2014	newstest2015	newstest2016	newstest2017	average
Vaswani et al. (2017)	28.4	-	-	-	
Shaw et al. (2018)	29.2	-	-	-	
our bitext	29.2	31.4	35.0	29.4	31.2
+ RTT APE	30.7	31.2	33.6	30.1	31.4
our NBT	33.5	34.4	38.3	32.5	34.7
+ RTT APE (bitext RTT)	32.5	32.7	35.2	31.3	32.9

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Is APE not working for larger domains?

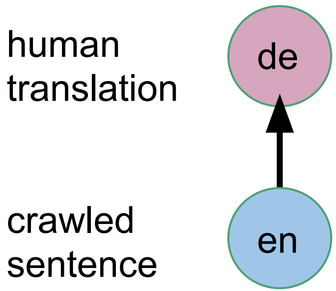
History Lesson: IWSLT 2011(?) Eval

- New Compound-Splitter for German that outperformed our old one by several BLEU on other non-IWSLT tasks
- It did not show any impact on IWSLT
    - WHY???

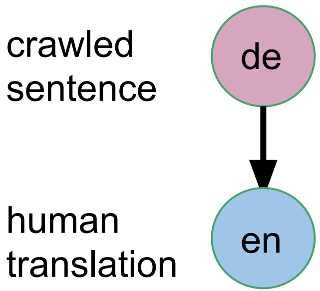
History Lesson: IWSLT 2011(?) Eval

- New Compound-Splitter for German that outperformed our old one by several BLEU on other non-IWSLT tasks
- It did not show any impact on IWSLT
- How is the test set constructed?

IWSLT test set



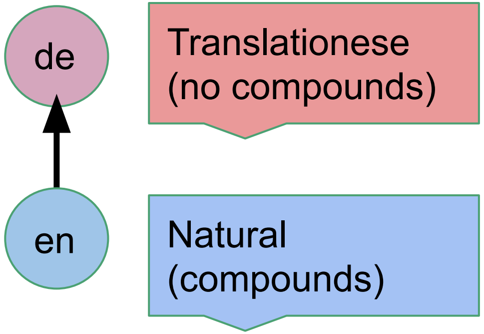
non-IWSLT test set



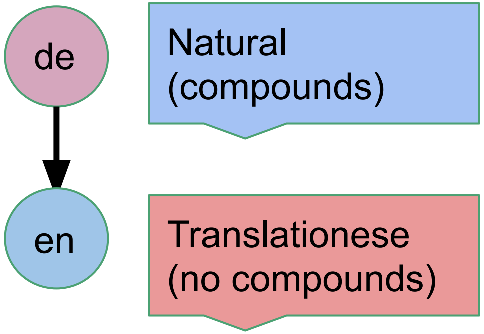
History Lesson: IWSLT 2011(?) Eval

- New Compound-Splitter for German that outperformed our old one by several BLEU on other non-IWSLT tasks
- It did not show any impact on IWSLT
- How is the test set constructed?

IWSLT test set

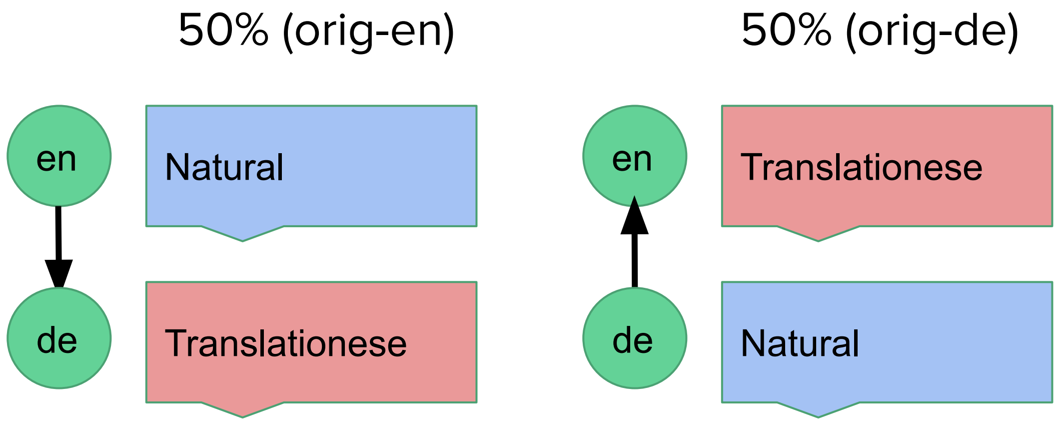


non-IWSLT test set



WMT test sets (since 2014)

- Each test set is



- Does our **APE system naturalize** text so that it **better matches reference** sentences that are **original in German** and not translationese?

BLEU by Original Language

	average	
	orig-de	orig-en
our bitext	27.7	33.1
+ RTT APE	33.3	29.8
our NBT	34.4	34.3
+ RTT APE	35.7	30.7

- orig-de: Input: Translationese, Reference: Natural
- orig-en: Input: Natural, Reference, Translationese

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

BLEU by Original Language

	average	
	orig-de	orig-en
our bitext	27.7	33.1
+ RTT APE	33.3	29.8
our NBT	34.4	34.3
+ RTT APE	35.7	30.7

- orig-de: Input: Translationese, Reference: Natural
- orig-en: Input: Natural, Reference, Translationese

➡️ **APE only works on the orig-de half of the test sets**

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Apply APE on de-orig Side

	newstest2014	newstest2015	newstest2016	newstest2017	average
Vaswani et al. (2017)	28.4	-	-	-	
Shaw et al. (2018)	29.2	-	-	-	
our bitext	29.2	31.4	35.0	29.4	31.2
+ RTT APE	30.7	31.2	33.6	30.1	31.4
+ RTT APE de-orig only	31.7	32.9	37.2	31.9	33.4
our NBT	33.5	34.4	38.3	32.5	34.7
+ RTT APE (bitext RTT)	32.5	32.7	35.2	31.3	32.9
+ de-orig only (bitext RTT)	34.0	34.5	38.7	33.2	35.1

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Results replicate across languages

	dev	test
Sennrich et al. (2016a)	-	28.8
our bitext	27.0	28.9
+ RTT APE	27.3	29.0
+ RTT APE only ro-orig	30.0	29.2

[WMT EnRo]

	newstest2014
our bitext	43.2
+ RTT APE	43.3
+ RTT APE only fr-orig	44.2
our NBT	45.3
+ RTT APE	44.6
+ RTT APE only fr-orig	46.1

[WMT EnFr]

Results on best WMT submissions

	Microsoft	Cambridge
WMT18 submission	48.7	47.2
+ APE only de-orig	49.5	47.7

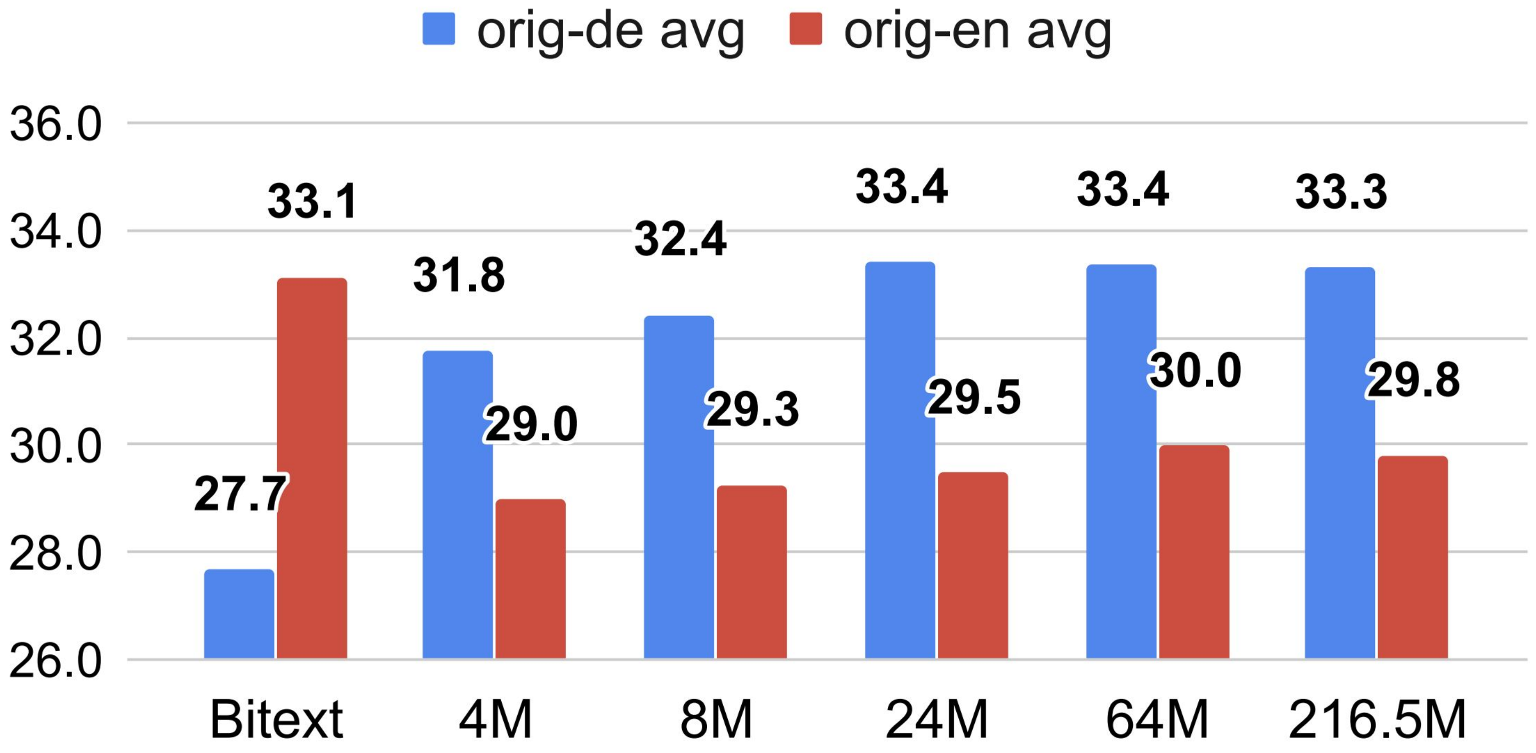
[WMT EnDe]

	QT21	Edinburgh
WMT16 submission	29.4	28.8
+ RTT APE only ro-orig	29.7	29.0

[WMT EnRo]

*best = best by BLEU

Different Training Data Sizes



But what about the other half?

Question

So far only run APE on sentences with **translationese input and natural reference** to reach maximum BLEU score!

Does human agree with the drop in performance for that half of the test set?

Human Evaluation

	newstest2016			
	fluency		accuracy	
	orig-de	orig-en	orig-de	orig-en
baseline bitext	4.65		95.6%	
+ RTT APE	4.77		98.4%	
our NBT	4.79		98.2%	
+ RTT APE	4.82		98.0%	
reference	4.85		98.6%	

- Are the BLEU gains actual quality gains? **yes!**

Human Evaluation

	newstest2016			
	fluency		accuracy	
	orig-de	orig-en	orig-de	orig-en
baseline bitext	4.65	4.49	95.6%	94.4%
+ RTT APE	4.77	4.59	98.4%	95.0%
our NBT	4.79	4.64	98.2%	95.8%
+ RTT APE	4.82	4.63	98.0%	96.2%
reference	4.85	4.67	98.6%	98.6%

-6.0 BLEU

-5.8 BLEU

- Are the BLEU gains actual quality gains? **yes!**
- Are the BLEU losses actual quality losses? **no!**
- Is only the fluency improving, but not the accuracy? **Both are improving!**

Implications on MT Evaluation

There are clear problems with translationese references:

- Human translators will introduce “translationese” biases, so models producing **more natural text may be penalized**
- This holds for any reference-based evaluation metric: BLEU, TER, …

But there are also problems with natural references:

- They do **not represent any real-world** translation task
- Translationese sources may be **much easier to translate**

Implications on APE Systems

There are clear problems with translationese references:

- Human translators will introduce “translationese” biases, so APE models (trained with synthetic data) producing **more natural text may be penalized**
- This holds for any reference-based evaluation metric: BLEU, TER, ...

But there are also problems with natural references:

- They do **not represent any real-world** translation task
- Translationese sources may be **much easier to translate**

Discussion

1. We encourage researchers to **split test sets** based by their original language and **report scores on both** subsets.
2. Were APE model that use synthetic data underestimated?
3. Can we generate a **MT system that produces natural translations** and overcomes the translationese biases of human translators?

Are Translationese Real?

- Koppel and Ordan (2011) train a high-accuracy **classifier** to distinguish **human-translated text from natural text** in the Europarl corpus.
- Well-known in professional translation world: both systematic **biases inherent to translated texts** (Baker, 1993; Selinker, 1972), as well as **biases** resulting specifically from interference **from the source text** (Toury, 1995).
- Similarly: **conflict between *Fidelity*** (the extent to which the translation is faithful to the source) and ***Transparency*** (the extent to which the translation appears to be a natural sentence in the target language)

Ablation

Iterative APE

Can we further improve/naturalize the MT output when we iteratively apply APE?

- Apply the same APE model on the already automatic post-edited output

	average	
	orig-de	orig-en
our bitext	27.7	33.1
+ APE	33.3	29.8
+ 2xAPE	33.2	29.1

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Reverse APE model

- Flip source and target of the APE training data and train an APE on (y, RTT(y)) sentence pairs.
- Goal is to translationese our output
- On original-en half of the test set:
    - Reverse APE outperforms APE on BLEU

	average	
	orig-de	orig-en
our bitext	27.7	33.1
+ RTT APE	33.3	29.8
+ Reverse APE	25.1	30.6
our NBT	34.4	34.3
+ RTT APE	35.7	30.7
+ Reverse APE	27.0	31.3

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Inside the Black Box of RTT

How much does RTT changes the translation output:

- BLEU = 40.9
- Unigram precision = 72.3%
- Bigram precision = 48.9%
- Trigram precision = 35.6%
- 4gram precision = 26.6%

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Inside the black box of RTT

What about the vocabularies of natural vs RTT sentences?

- Vocabulary size of natural text = 33,814
- Vocabulary size of RTT = 29,635

- Vocabulary size of NMT+APE output = 31,471
- Vocabulary size of NMT output = 30,540

➡️ In this setup: larger vocabulary = higher performance

[WMT18 English->German: APE trained on ~216M newscrawl sentences.]

Accuracy Examples

source	Using a **club**, they **beat** the victim in the face and upper leg.
NBT	Mit einem **Club schlagen** sie das Opfer in Gesicht und Oberschenkel.
+ RTT APE	Mit einem **Schlagstock schlugen** sie dem Opfer ins Gesicht und in den Oberschenkel.
source	Müller put another one in with a **with a penalty**.
NBT	Müller setzte einen weiteren **mit einer Strafe** ein.
+ RTT APE	Müller netzte einen weiteren **per Elfmeter** ein.
source	Obama **receives** Netanyahu
NBT	Obama **erhält** Netanjahu
+ RTT APE	Obama **empfängt** Netanjahu
source	At least one Bayern fan was **taken injured from the stadium**.
NBT	Mindestens ein Bayern-Fan wurde **vom Stadion verletzt**.
+ RTT APE	Mindestens ein Bayern-Fan wurde **verletzt aus dem Stadion gebracht**.
source	The archaeologists **made a find in the third construction phase** of the Rhein Boulevard.
NBT	Die Archäologen **haben in der dritten Bauphase** des Rheinboulevards **gefunden**.
+ RTT APE	Die Archäologen **sind im dritten Bauabschnitt** des Rheinboulevards **fündig geworden**.

Summary

1. An APE system trained on synthetic dataset **does improve the quality** of the MT output if the underlying **NMT system is a black box**
2. It **does not improve the quality** if we augment the NMT training data with **NBT**
3. Using **translationese as references** is not perfect as it **penalizes** output that is more **natural** (this is in particular important for APE systems trained on synthetic data)
4. Are APE models actually better than we thought?

Thanks!