Neural Machine Translation of Text from Non-Native Speakers

Antonios Anastasopoulos* Alison Lui*
Toan Q. Nguyen and David Chiang
aanastas@andrew.cmu.edu

NAACL 2019

Carnegie Mellon University

The sad truth

Non-native English speakers outnumber native ones by:

3:1

[Crystal, 2003]

…and we make "mistakes"

Yet, can our models handle non-standard input?

Do our datasets cover non-standard English?

Machine Translation and Noise

Yonatan2 : "Synthetic and Natural Noise both break NMT"

If you can raed tihs, you msut be ralely samrt

…but the MT system is not so it struggles to translate it!

Machine Translation and Noise

eng: This is not such a big deal due to the restart system, but is enough to spoil the fun.

Machine Translation and Noise

eng: This is not such a big deal due to the restart system, but is enough to spoil the fun.

spa: No se trata de mucho debido as sistema de reinicio, sino que basta con echar por tierra el fun.

Machine Translation and Noise

eng: This is not such a big deal due to the restart system, but is enough to spoil the fun.

↳ spa: No se trata de mucho debido as sistema de reinicio, sino que basta con echar por tierra el fun.

eng: This is not such a big deal due to the restart system, but is enough to spoil of the fun.

Machine Translation and Noise

eng: This is not such a big deal due to the restart system, but is enough to spoil the fun.

↳ spa: No se trata de mucho debido as sistema de reinicio, sino que basta con echar por tierra el fun.

eng: This is not such a big deal due to the restart system, but is enough to spoil of the fun.

↳ spa: ~~No se trata de~~Esto no es mucho debido al sistema de reinicio, sino que basta con ~~echar~~ echarnos por tierra ~~el fun~~.

Working with Realistic Noise

Grammar Error Correction Corpora

English from non-native speakers with corrections.

Create confusion sets for some error types:

p(error | correct)

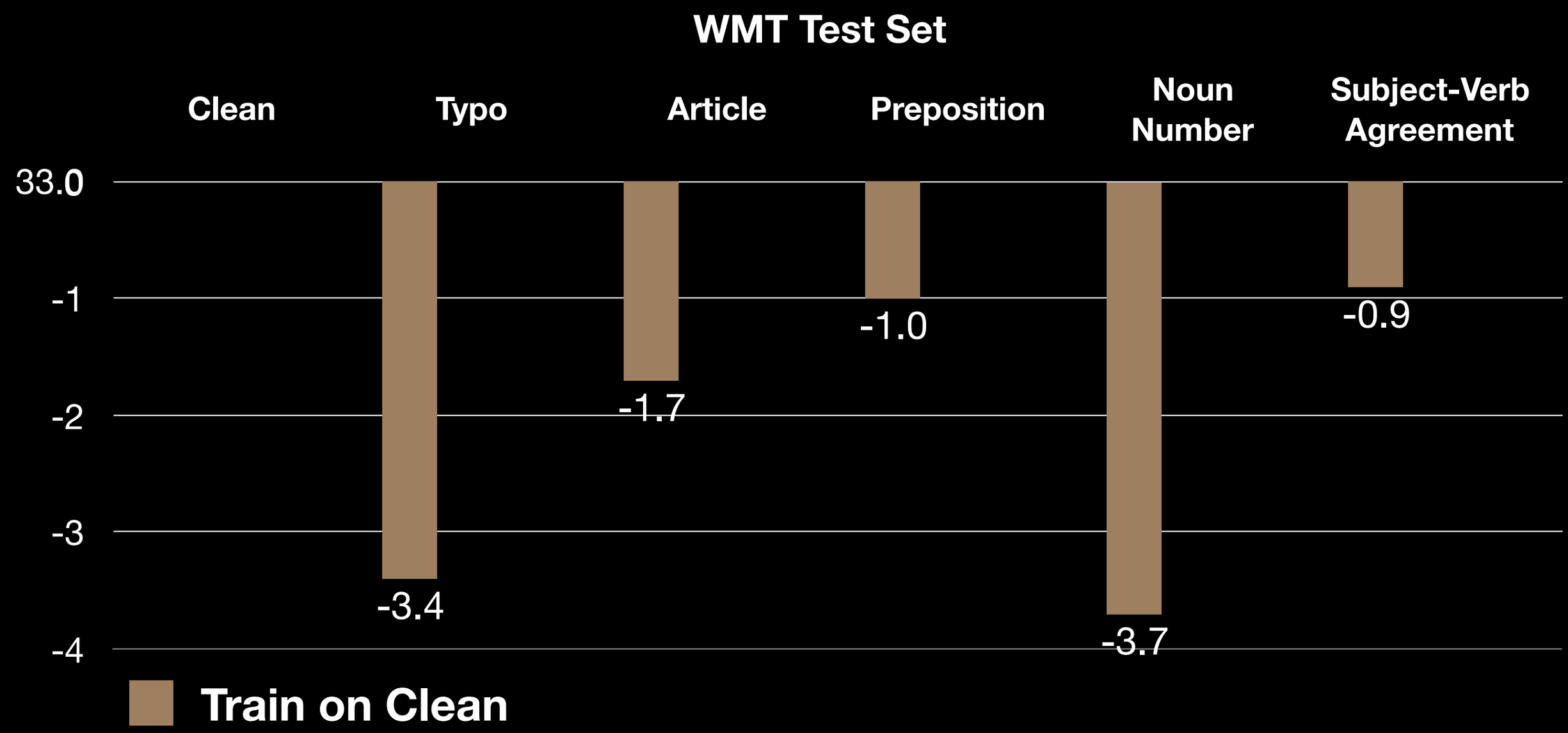
	a	an	the	Ø
a		0.01	0.18	0.8
an	0.06		0.14	0.78
the	0.02	0.0		0.96
Ø	0.08	0.01	0.9	

Working with Realistic Noise

We added a **single** error on a WMT English test set

```
clean:     Yet the debit market is breaking records today.

typo:      Yet the debit maret is breaking records today.

article:   Yet the debit market is breaking records the today.

prep:      Yet the debit market is breaking records in today.

noun num:  Yet the debit market is breaking record today.

sva:       Yet the debit market are breaking records today.
```

Testing on Grammatical Noise



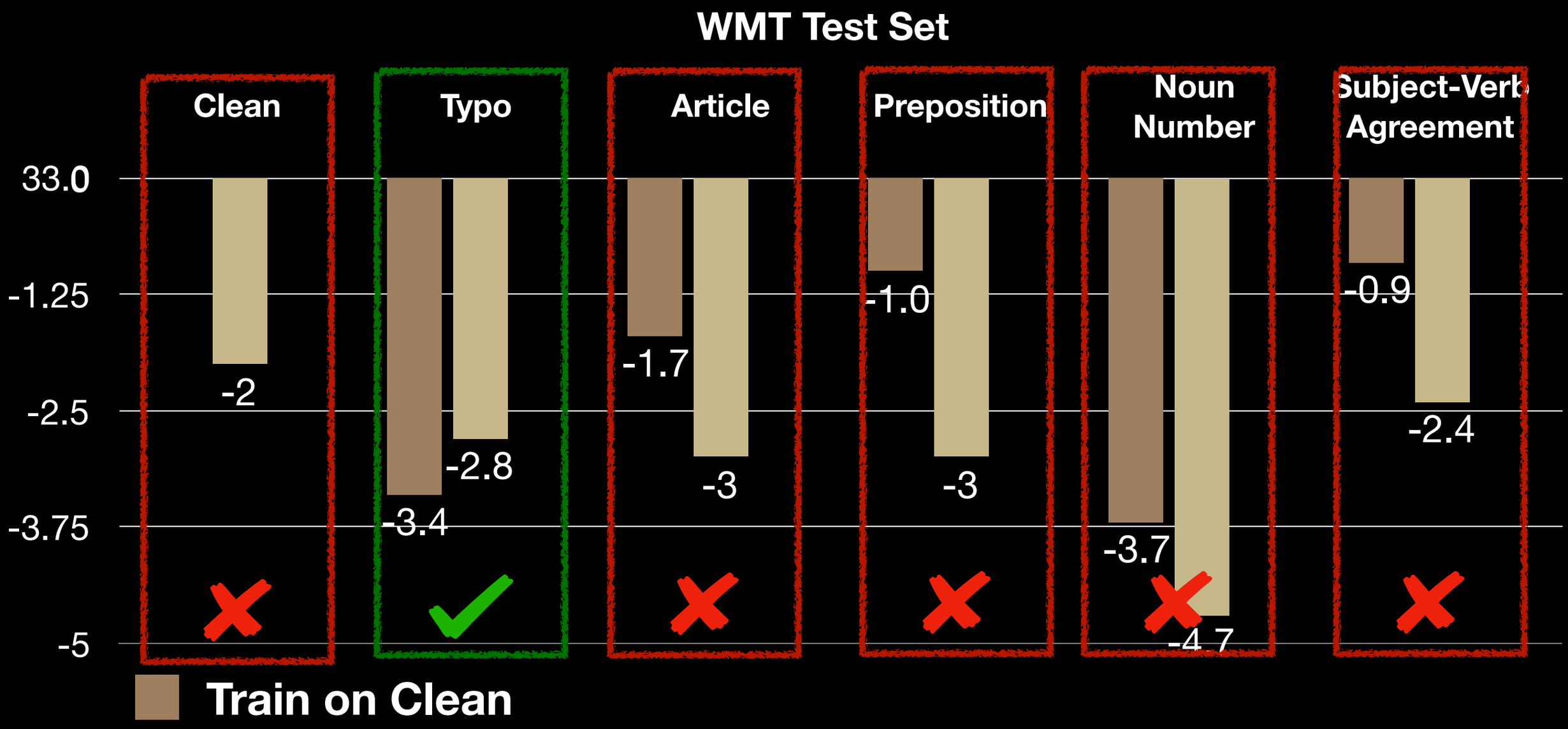
Solution

Add noise to the training data…

…and train with the synthesized data.

First, add a single type of error.

Testing on Grammatical Noise



1. **Improvements only on the respective test set**
2. **Worse performance on clean data**

Solution

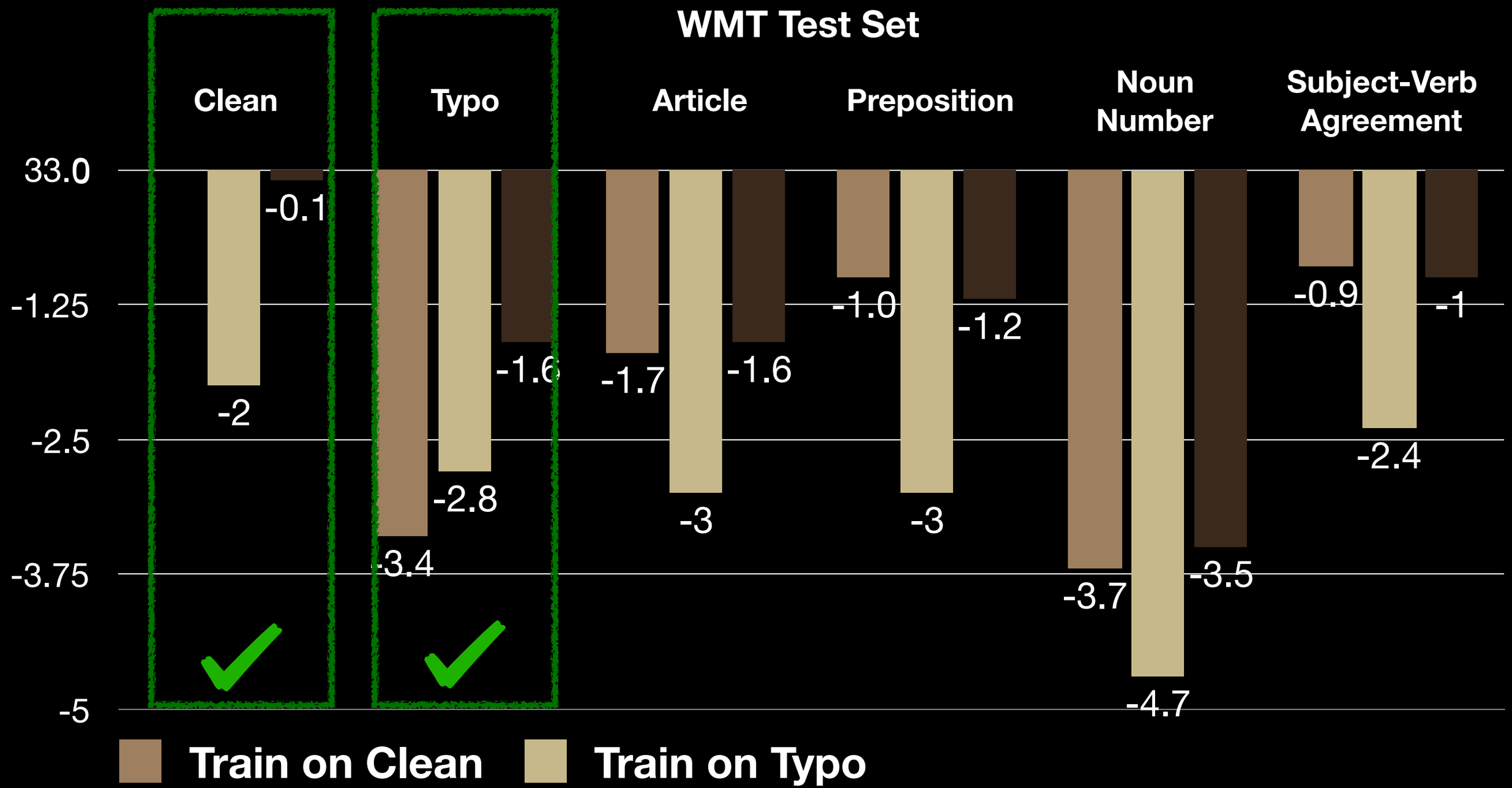
Add noise to the training data…

~~…and train with the synthesized data.~~

…and train on clean **and** synthesized data (concatenated).

First, add a single type of error.

Testing on Grammatical Noise



Solution

Add noise to the training data…

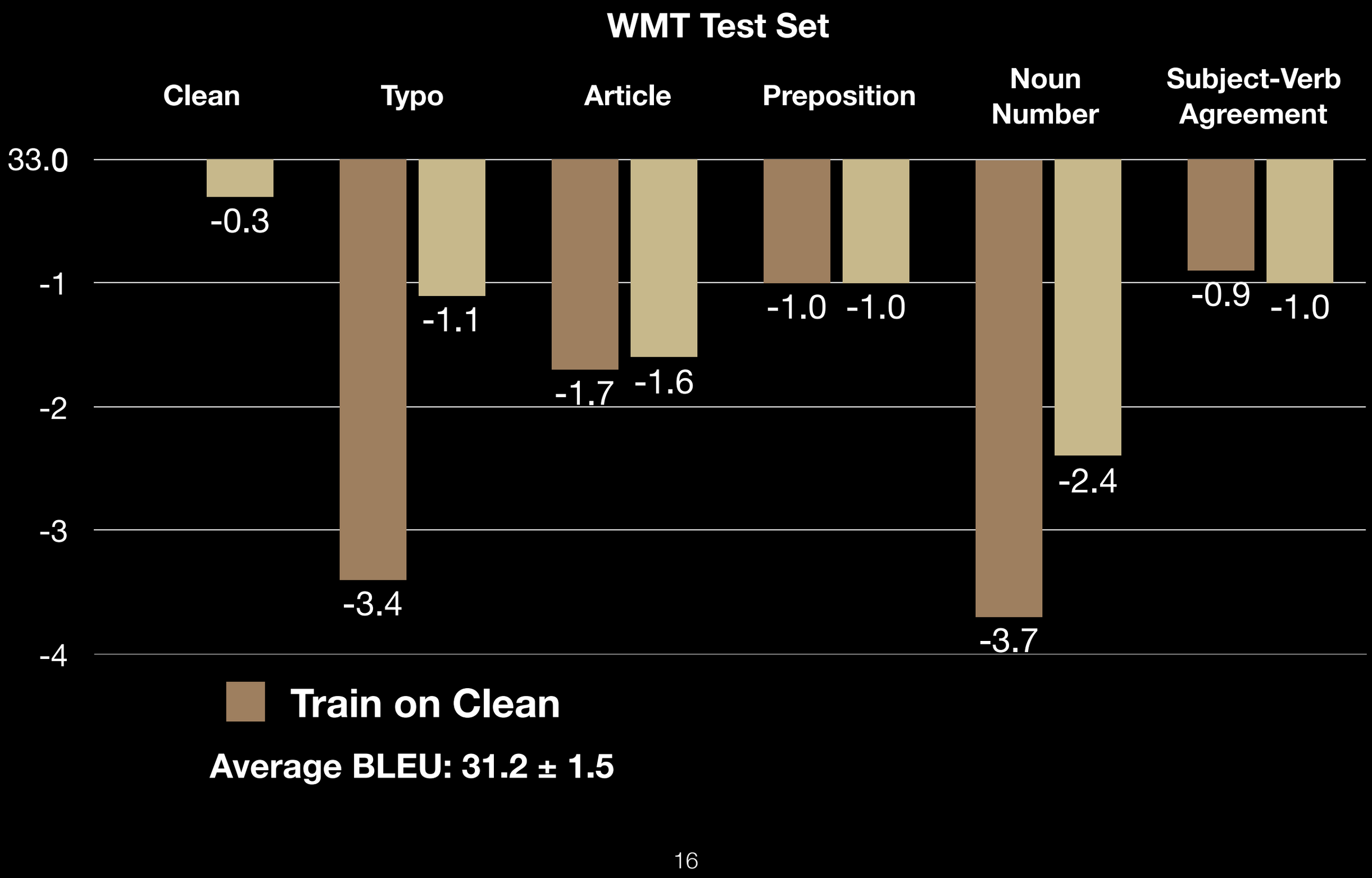
~~…and train with the synthesized data.~~

…and train on both clean and synthesized data.

~~First, add a single type of error.~~

… adding all types of errors.

Testing on Grammatical Noise



Working with *Real* Noise

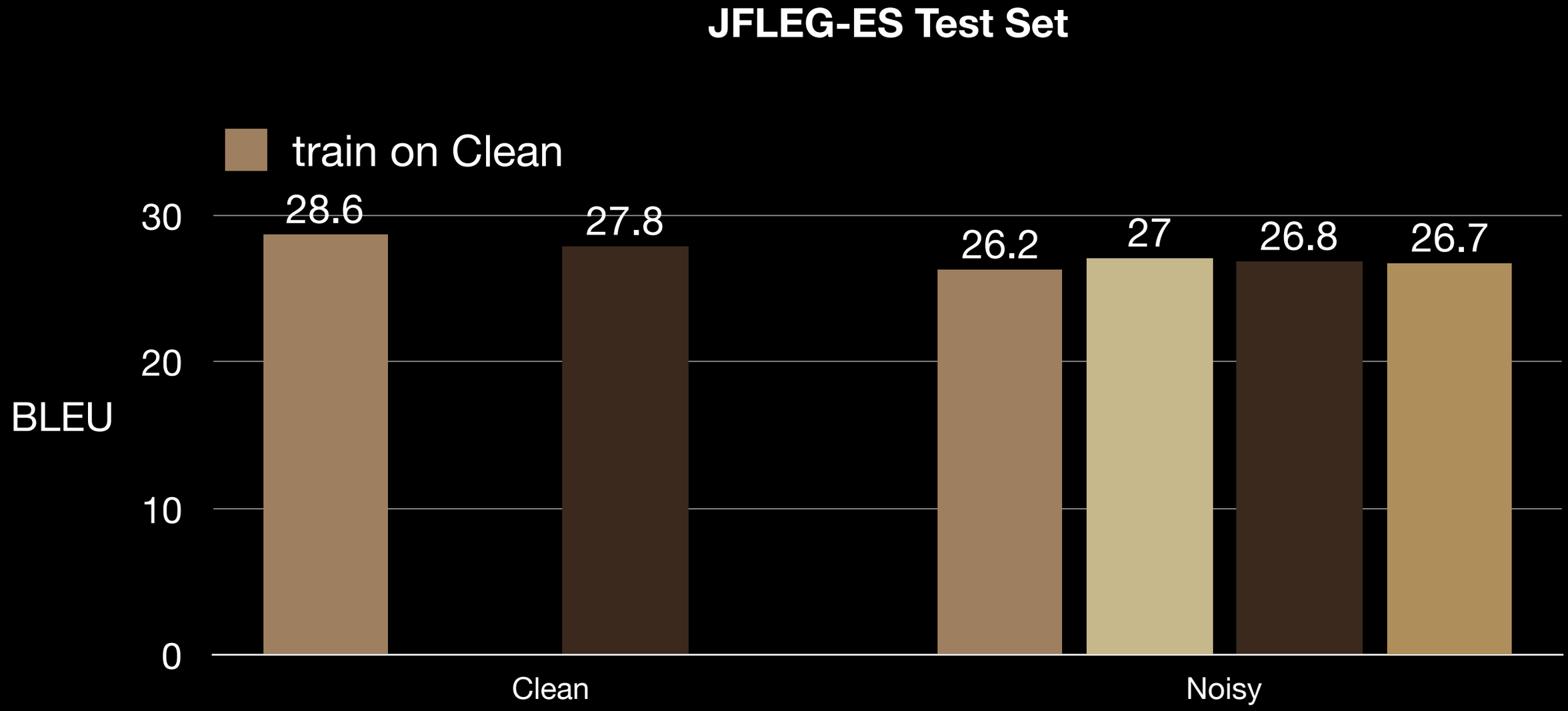
JHU Fluency-Extended GUG Corpus (JFLEG)

Prompt:

Please translate the following sentences.
Note that some sentences will have grammatical errors or typos in English. Don't try to translate the sentences word for word (e.g. replicate the error in Spanish).

JFLEG-ES

Testing on Real Grammatical Noise



- **2.4 BLEU loss due to noise**
- **What if we used a GEC system as a pre-processing step?**
- **Small loss on clean, but improvement on noisy**

Summary

We need to account for non-native speakers and language learners

We need to create appropriate datasets. We provide

1. train, dev, and test WMT datasets
   with realistic noise on the English side on
   8 language pairs: {cs,de,et,fi,ru,tr,zh}-en
2. JFLEG-ES: an evaluation set (dev+test) with professional translations in Spanish

Available at: **https://bitbucket.org/antonis/nmt-grammar-noise/**

Towards robustness:

combine synthesized noisy data with clean ones