On the Impact of Various Types of Noise on Neural Machine Translation

Huda Khayrallah & Philipp Koehn
Center for Language & Speech Processing, Computer Science Department
Johns Hopkins University
{huda, phi}@jhu.edu

Abstract

We examine how various types of noise in the parallel training data impact the quality of neural machine translation systems. We create five types of artificial noise and analyze how they degrade performance in neural and statistical machine translation. We find that neural models are generally more harmed by noise than statistical models. For one especially egregious type of noise they learn to just copy the input sentence.

Motivation

- MT systems (especially NMT) are data hungry
- Is all data helpful?
- Adding noisy web-crawl hurts **NMT** and helps **SMT**

	NMT	SMT
WMT17	27.2	24.0
+ noisy corpus	17.3 (-9.9)	25.2 (+1.2)

What kind of noise exists in the noisy corpus?

Type of noise	
Okay	23%
Misaligned sentences	41%
3rd Language	3%
Both English	10%
Both German	10%
Untranslated	4%
Short Segments ≤ 2	1%
Short Segments 3 - 5	5%
Other Text	2%

Experiments

- SMT (Moses) and NMT (Marian)
- WMT 2017 DE → EN baseline
- Web-crawled data from paracrawl.eu

Results

BLEU scores after adding different amounts of noise (ratio of original clean corpus)

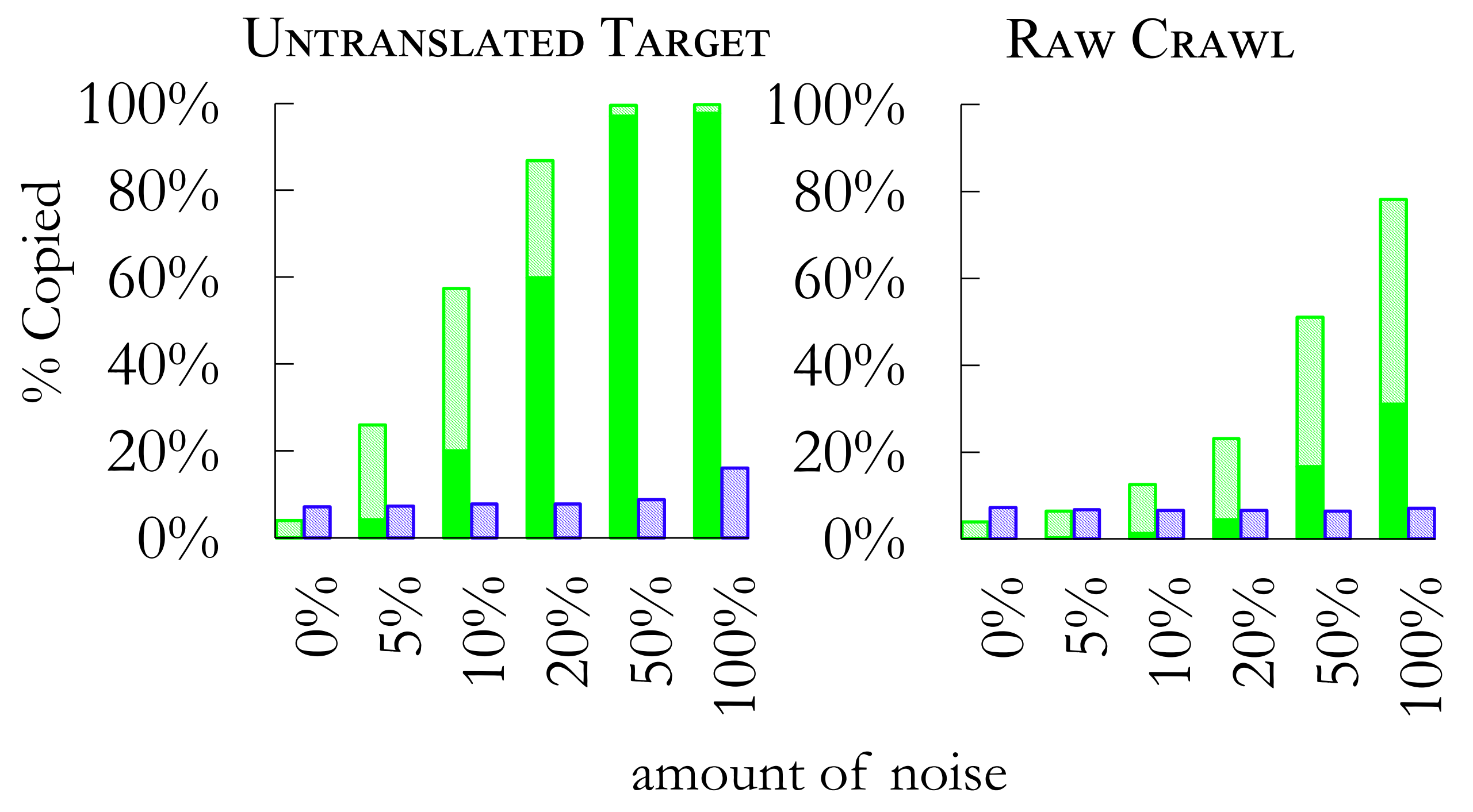
NMT (left green bars) is harmed more than **SMT (right blue bars)**

	5%		10%		20%		50%		100%	
Misaligned sentences	26.5 -0.7	24.0 -0.0	26.5 -0.7	24.0 -0.0	26.3 -0.9	23.9 -0.1	26.1 -1.1	23.9 -0.1	25.3 -1.9	23.4 -0.6
Misordered words (source)	26.9 -0.3	24.0 -0.0	26.6 -0.6	23.6 -0.4	26.4 -0.8	23.9 -0.1	26.6 -0.6	23.6 -0.4	25.5 -1.7	23.7 -0.3
Misordered words (target)	27.0 -0.2	24.0 -0.0	26.8 -0.4	24.0 -0.0	26.4 -0.8	23.4 -0.6	26.7 -0.5	23.2 -0.8	26.1 -1.1	22.9 -1.1
Wrong language (source)	26.9 -0.3	24.0 -0.0	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.8 -0.2
Wrong language (target)	26.7 -0.5	24.0 -0.0	26.6 -0.6	23.9 -0.1	26.7 -0.5	23.8 -0.2	26.2 -1.0	23.5 -0.5	25.0 -2.2	23.4 -0.6
Untranslated (source)	27.2 -0.0	23.9 -0.1	27.0 -0.2	23.9 -0.1	26.7 -0.5	23.6 -0.4	26.8 -0.4	23.7 -0.3	26.9 -0.3	23.5 -0.5
Untranslated (target)	17.6 -9.8	23.8 -0.2	11.2 -16.0	23.9 -0.1	5.6 -21.6	23.8 -0.2	3.2 -24.0	23.4 -0.6	3.2 -24.0	21.1 -2.9
Short segments (max 2)	27.1 -0.1	24.1 +0.1	26.5 -0.7	23.9 -0.1	26.7 -0.5	23.8 -0.2				
Short segments (max 5)	27.8 +0.6	24.2 +0.2	27.6 +0.4	24.5 +0.5	28.0 +0.8	24.5 +0.5	26.6 -0.6	24.2 +0.2		
Raw crawl data	27.4 +0.2	24.2 +0.2	26.6 -0.6	24.2 +0.2	24.7 -2.5	24.4 +0.4	20.9 -6.3	24.8 +0.8	17.3 -9.9	25.2 +1.2

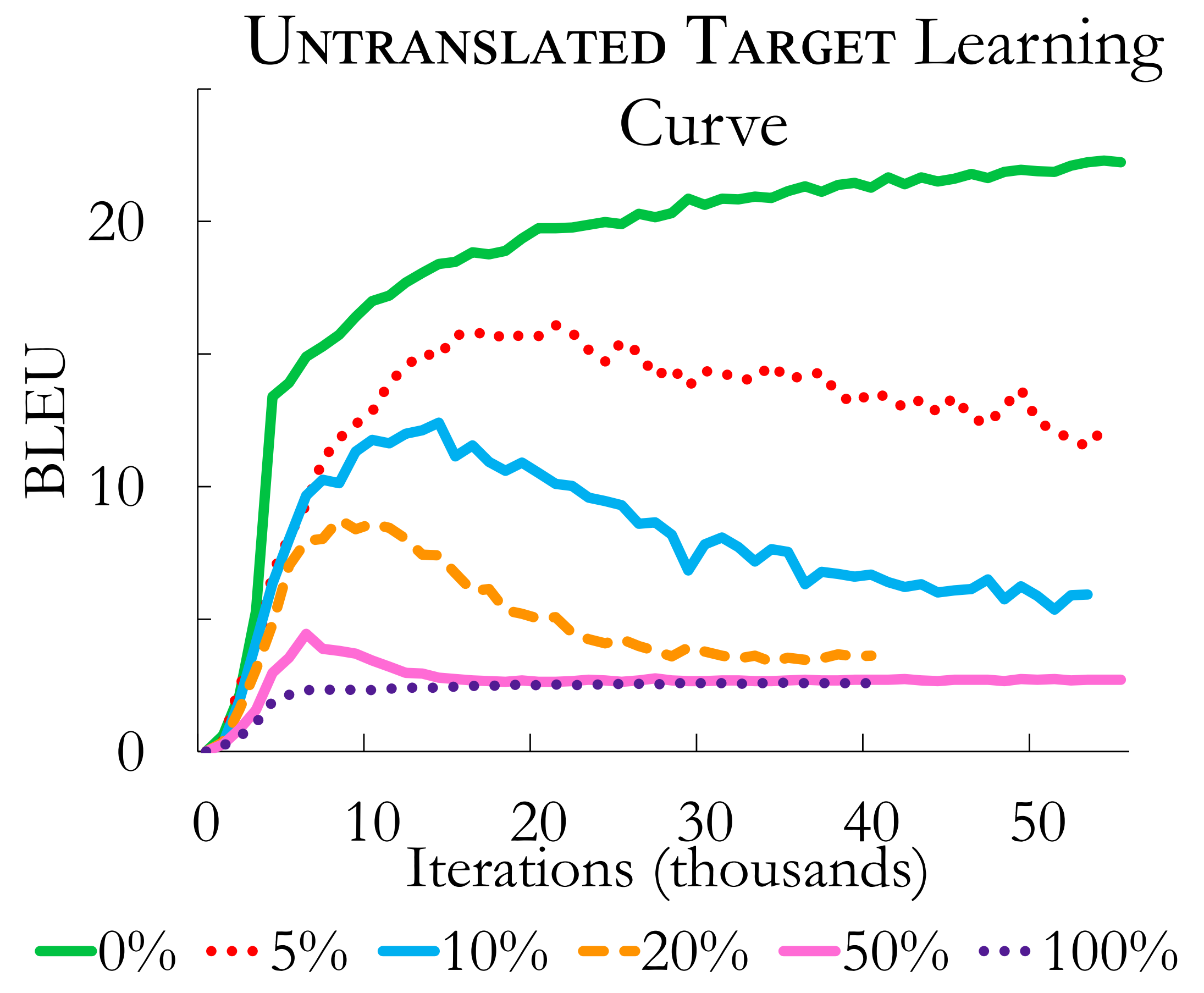
Analysis

What goes wrong with Untranslated Target?

- NMT learns to copy input, causing degradation.



- Copied sentences in the Untranslated Target & Raw Crawl experiments for **NMT (left green bars)**, and **SMT (right blue bars)**.
- Sentences that are exact matches to the source are the solid bars, sentences that are more similar to the source than the target are the shaded bars.



- Performance of the systems trained on noisy corpora begin to improve, before over-fitting to the copy portion of the training set.