MTNT: A Testbed for Machine Translation of Noisy Text

Paul Michel and **Graham Neubig**
Language Technologies Institute
Carnegie Mellon University
{pmichel1,gneubig}cs.cmu.edu

The topic of the robustness of Machine Translation (MT) systems to noise in text is receiving an increasing amount of interest. Noise in social media text is a known issue that has been investigated in a variety of previous work (*e.g.* Baldwin et al. (2013)). Most recently, Belinkov and Bisk (2018) have focused on the difficulties that character based Neural MT models have translating text with character-level noise within individual words. However, word-by-word replacement or scrambling of characters doesn't cover all the idiosyncrasies of Internet linguistics.

Despite the obvious utility of creating noise-robust MT systems, and the scientific challenges contained therein, there exists no standard open benchmark for researchers and developers of MT systems to test the robustness of their models to these and other phenomena found in noisy data on the Internet.

We introduce Machine Translation of Noisy Text (MTNT), a new dataset to test robustness of MT systems against noisy social media text. The dataset contains naturally created noisy source sentences with professionally created translations both in a pair of typologically close languages –English and French– and distant languages –English and Japanese. We collect noisy comments from the Reddit online discussion website in English, French and Japanese and ask professional translators to translate to and from English, resulting in 1000 test samples and from 6k to 36k training samples in four language pairs (en-fr, fr-en, en-ja and ja-en). In addition, we release additional small monolingual corpora in those 3 languages to both provide data for semi-supervised adaptation approaches as well as noisy Language Modeling experiments.

We quantitatively examine the types of noise included in this dataset by running the test sets through widely used spell-checking software[1] as well as checking for emojis and profanities. We compare those numbers to existing test sets for NMT and show that our dataset contains more noise (see table 1 for results on the English test set).

	newstest2014	MTNT
Spelling errors	0.21	2.18
Grammar errors	0.19	0.56
Emojis	0.00	0.29
Profanities	0.03	0.24

Table 1: Occurrences (per 100 tokens) of noisy phenomena.

We show that BLEU scores from models trained on traditional MT training corpora are much lower on MTNT than on common test sets from 19% to 79% lower depending on the language pair).

This indicates that this dataset can provide an attractive testbed for methods tailored to handling noisy text in MT. [2]

We intend this contribution to provide a standard benchmark for robustness to noise in MT and foster research on models, dataset and evaluation metrics tailored for this specific problem.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ICLR*.

[1] Grammarly for en and Microsoft Word for fr and ja

[2] The data will be made publicly available at http://redacted upon publication.