Translating User Generated Content: Challenges and Baselines

José Carlos Rosales[1] Djamé Seddah[1] Guillaume Wisniewski[2]
[1] *Inria Paris, Université Paris-Diderot*
[2] *LIMSI - CNRS, Université Paris-Sud*

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ambassade de France en Irlande

Executive summary

Main aspects:

- Description and statistics of two UGC parallel corpora for Machine Translation (`MTNT` [Michel & Neubig, 2018] and `CrapBank`).
- Quantifying noise/*domainess* on UGC automatically: % OOVs, perplexity, bits-per-character and KL divergence of 3-grams distribution.
- Statistical Machine Translation and Neural Machine Translation benchmarks for assessing UGC's difficulties.

Conclusions:

- Robustness to UGC comparison between phrase-based SMT and attentional decoding NMT.

User Generated Content

- Corpora produced by users (blogs, forums, social media, commentaries) without any restriction.
- Out-of-vocabulary are the main translation issues: high quantity of typos, grammar/vocabulary errors (mainly due to french homophonies), abbreviations, jargon, named entities.
  ⇒ New fr-en UGC parallel corpus `CrapBank`: for every new corpus there are new challenges

Examples:

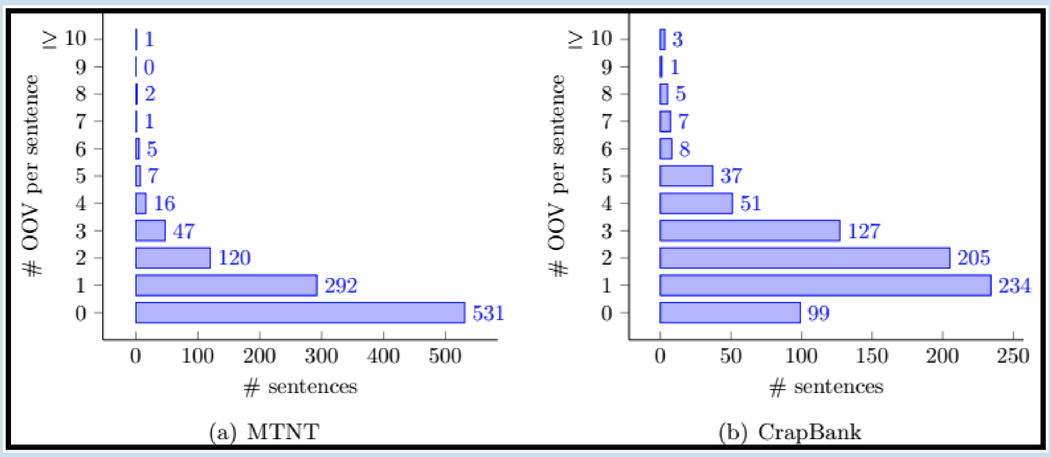
- **CrapBank UGC**:
  - **FR (src):** pour Victoire de la Musique pour Vanessa Paradis on l'aime trait fort il faux quel a la Victoire elle doit gagne je saurai trop content pour elle.
  - **EN (ref):** for Victoire de la Musique to Vanessa Paradis we love her very much she has to have a Victoire she has to win I would be so happy for her.
  - **EN (conventional output):** for Victory of the Music to Vanessa Paradis we love treating her strongly it fake which Victory for her she has win I would know too happy for her.

Corpora statistics

Corpus	Subset	#sentences	#words	avg. sent.len	TTR
MTNT*	train	19,161	798,809	41.7	0.10
	test	1,022	20,169	19.7	0.34
CrapBank*	train	777	15,960	20.5	0.37
	test	777	13,680	17.6	0.32
WMT (Training Corpus)	train	2.2×10^6	64.2×10^6	29.7	0
News Test (2015)	test	3,003	82,775	27.6	0.23
NewsDiscussTest (2014)	test	1,500	30,083	20.0	0.30

UGC's noise quantification

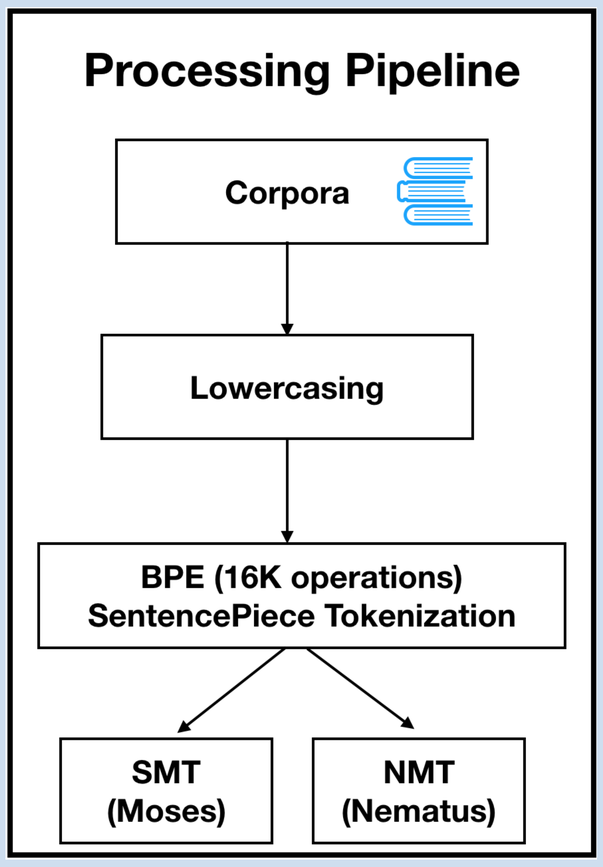
Corpus	Subset	Perplexity	BPC	3-grams KL-div.	% OOV
MTNT	train (fr)	353.07	8.46	0.85	9.66%
	test (fr)	363.14	8.50	0.79	10.14%
CrapBank	train (fr)	1034.76	10.01	1.92	21.78%
	test (fr)	1030.23	10.01	1.86	20.87%
WMT (Training Corpus)	train(fr)	7.34	2.88	0	0%
	train (en)	8.19	3.04	0	0%
News Test (2015)	test (fr)	102.92	6.69	0.46	3.66%
NewsDiscussion Test (2014)	test (fr)	114.90	6.84	0.48	3.89%



(a) MTNT (b) CrapBank

Model and Protocol

- Phrase-based SMT:
  - 5-grams Kneser-Ney Language Model (`lmplz`).
  - Word alignments computed with `GIZA++`.
- Neural MT:
  - Two 512 LSTM layers on the encoder and decoder sides.
  - Conventional attentional decoder [Bahdanau et al. 2014].



Results

Not tuned baselines:

Corpus	NMT	SMT
NewsTest	28.93	24.55
NewsDiscTest	30.76	27.56
CrapBank Test	13.43†	23.68
MTNT Test	23.27	24.03

Tuned baselines (with `MTNT` Train set):

Corpus	NMT	SMT
NewsTest	———	25.72
NewsDiscTest	———	28.13
CrapBank Test	19.04†	28.95
MTNT Test	30.29	29.86

- The metric used is `SacreBleu` [Post, 2018].
- Scores marked with † are reproduced by using the author's implementation.

Acknowledgements

The Embassy of France in Ireland for funding our participation in the conference.
This work have been partialy funded by l'Agence Nationale de la Recherche (projet ParSiTi, ANR-16-CE33-0021).