Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks Supplementary Material

Jörg Tiedemann and **Yves Scherrer**
Department of Digital Humanities / HELDIG
University of Helsinki

1 Translation

Table 1 summarizes the BLEU scores when testing on heldout data from the in-domain corpus (Bible) and the out-of-domain corpus (Tatoeba). For the former we used heldout data from the English standard Bible and the New-World Bible for French. For Tatoeba, we created a multi-reference test corpus from the English-French translations in the database that includes 1,068 English sentences with a total of 7,998 translations into French.

Training languages	**Bible** BLEU	Δ	**Tatoeba** BLEU	Δ
English–French	21.29		15.62	
+ Afrikaans	21.14	-0.15	16.49	**0.87**
+ Albanian	21.22	-0.07	15.82	**0.20**
+ Breton	20.91	-0.38	15.43	-0.19
+ German	20.77	-0.52	14.63	-0.99
+ Greek	20.87	-0.42	15.43	-0.19
+ Frisian	21.59	**0.30**	15.52	-0.10
+ Hindi	21.47	**0.18**	15.07	-0.55
+ Italian	21.40	**0.11**	16.48	**0.86**
+ Dutch	21.18	-0.11	16.30	**0.68**
+ Ossetian	20.84	-0.45	17.11	**1.49**
+ Polish	21.05	-0.24	17.18	**1.56**
+ Russian	21.00	-0.29	15.49	-0.13
+ Slovene	21.40	**0.11**	16.30	**0.68**
+ Spanish	20.81	-0.48	15.11	-0.51
+ Serbian	21.44	**0.15**	17.19	**1.57**
+ Swedish	20.64	-0.65	16.85	**1.23**
average	21.11	-0.18	16.03	**0.41**

Table 1: English to French translation quality in terms of BLEU scores, using the in-domain Bible test set (left half, single reference) and the out-of-domain Tatoeba test set (right half, multiple references). The columns marked with Δ show the absolute BLEU score difference compared to the baseline English–French model; improvements are highligted in bold face.

In-domain translation with multilingual models is on par with bilingual ones and ot-of-domain models show gains in the majority of the cases up to 1.57 BLEU points.

2 Paraphrase Generation

Table 2 lists the percentage of identical copies produced when using multilingual NMT models for paraphrase generation. In the comparison we discard punctuations and compare lowercased strings. Adding English-to-English paraphrased training data significantly increases the percentage.

Model	**Bible**	**Tatoeba**
English–French	0.0%	0.7%
+ Afrikaans	0.9%	4.8%
+ Albanian	0.7%	3.4%
+ Breton	0.0%	1.1%
+ German	1.4%	4.9%
+ Greek	1.1%	5.2%
+ Frisian	0.7%	4.3%
+ Hindi	0.9%	4.2%
+ Italian	1.2%	5.0%
+ Dutch	1.1%	5.1%
+ Ossetian	0.6%	3.5%
+ Polish	0.4%	2.8%
+ Russian	1.4%	4.7%
+ Slovene	0.6%	3.2%
+ Spanish	1.1%	5.5%
+ Serbian	0.5%	3.3%
+ Swedish	1.2%	4.9%
+ All	0.8%	2.0%
+ English–English	71.6%	70.0%

Table 2: Percentages of identical source and generated target sentences. Multilingual models produce significantly less copies of the input compared to the supervised paraphrase model trained on pairs of English Bible variants (last line).