A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German—English Machine Translation Output

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, Hans Uszkoreit

German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab, Berlin

This work has received funding by the German Research Foundation (DFG) through the project "Analyse und automatische Abschätzung der Qualität maschinell generierter Text", and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear.

Test Suite Creation Process

- around 10,000 test items to evaluate German <> English MT outputs
- 13/14 categories per language direction, divided into more than 100 linguistic phenomena each, with min. 20 test items per phenomenon
- set of rules (fixed strings and regex) for the semi-automatic evaluation
- half of the test items and rules available at https://github.com/DFKI-NLP/mt-testsuite

Test Suite Application Process

- the MT-TestSuite was submitted to the WMT 2018-2021 (De-En) and 2021 (En-De)
- translations from systems of the news translation task were evaluated with the test suite

Application Example: WMT Test Suite Track

- the MT-TestSuite was submitted to the WMT 2018-2021 (De-En) and 2021 (En-De)
- translations from systems of the news translation task were evaluated with the test suite

Further Application Examples

- WMT metrics track
- Domain-specific test suite
- Quality estimation
- Portuguese—English test suite

Future Work

- automatic test item creator for test set expansion
- new publicly available evaluation tool
- include more language pairs, currently we are working on extending to English–Russian