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

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

MT-TestSuite

- 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 Creation Process

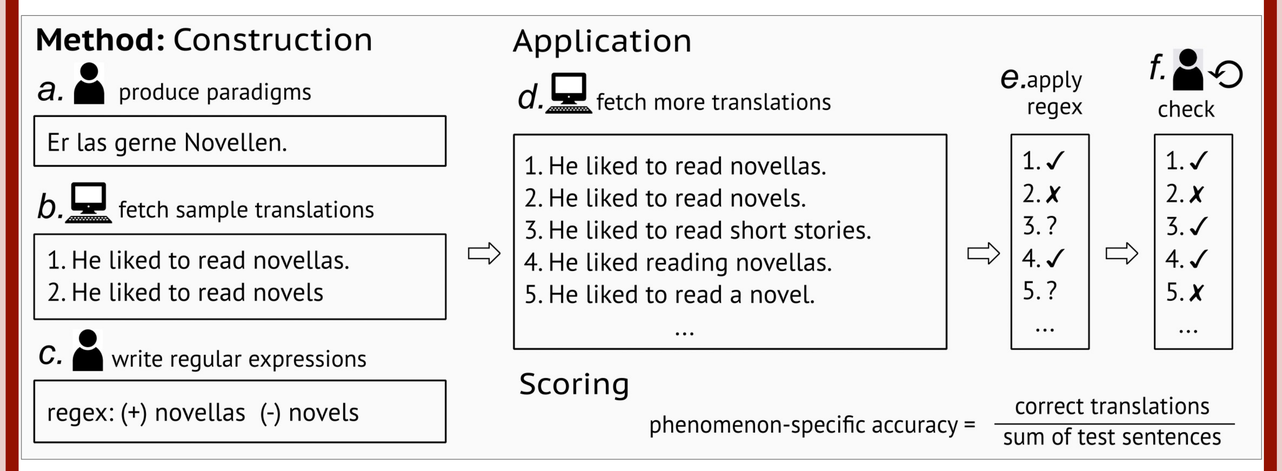

Figure 1: Example of the preparation of the test suite for one test item.

Test Suite Application Process

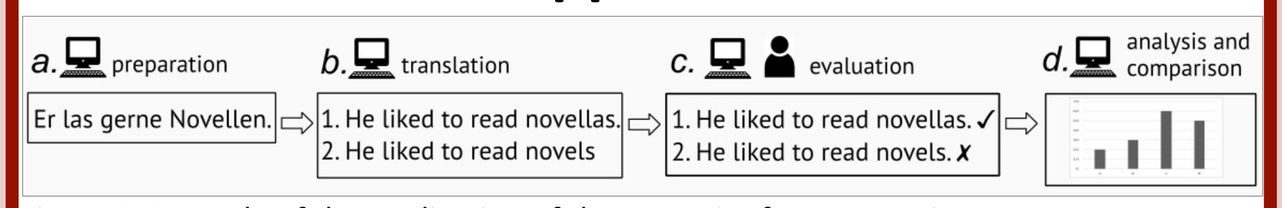

Figure 2: Example of the application of the test suite for one test item.

Test Suite Examples

	Collocation	False Friends	Modal negated preterite
Test item	Simon trinkt am liebsten *lieblichen Wein.*	Dieser Autor schreibt hauptsächlich *Novellen*.	Ihr *durftet nicht lesen.*
WMT 18	Simon prefers to drink *adorable wine.*	This author writes mainly *Novellen*.	You *are not allowed to read*.
WMT 19	Simon loves to drink *lovely wine.*	**This author writes mainly *novellas*.**	You *are not allowed to read.*
WMT 20	Simon prefers to drink *lovely wine.*	**This author writes mainly *novellas*.**	You *don‘t read.*
WMT 21	**Simon prefers to drink *sweet wine*.**	This author mainly writes *novels*.	**You *were not allowed to read.***

Table 1: Output examples from 2018 to 2021. *phenomenon,* **correct**

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

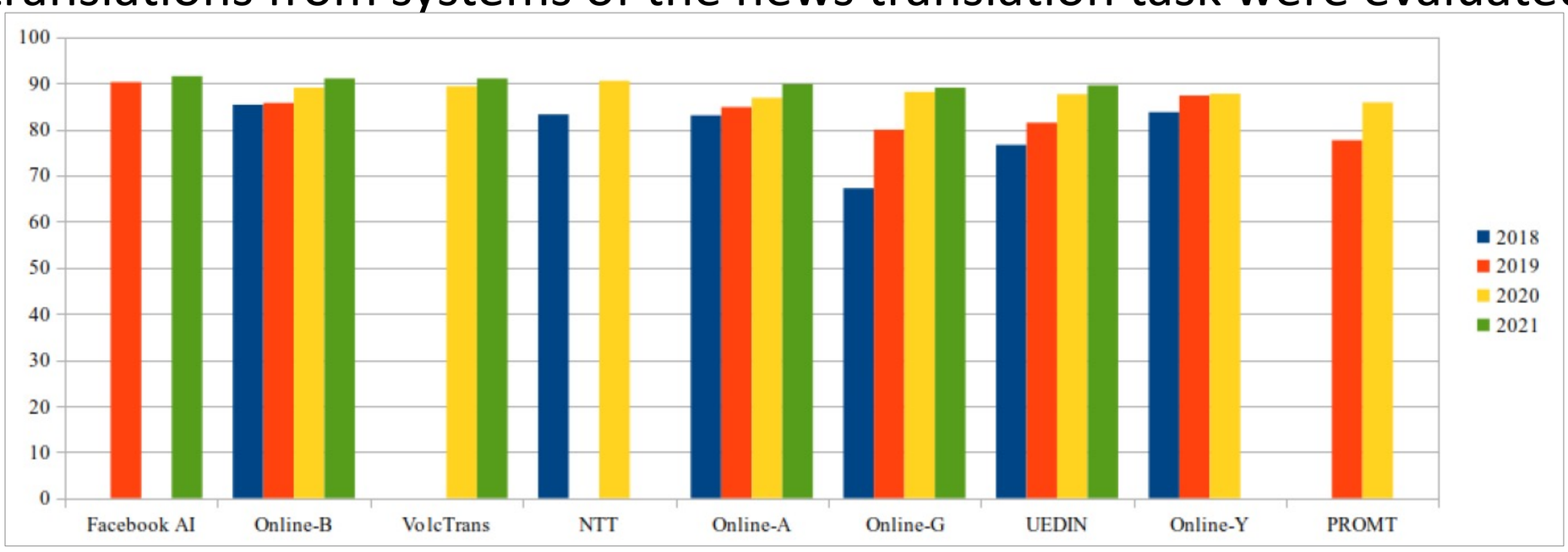

Figure 3: System improvements (accuracy macro-average on the test suite from 2018 to 2021 for German to English, including the systems that appear at least once in the past two years.

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

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.

Vivien Macketanz
Speech and Language Technology
Vivien.Macketanz@dfki.de