Fine-grained evaluation of Quality Estimation for Machine translation based on a linguistically-motivated Test Suite

Eleftherios Avramidis[*], Vivien Macketanz[*], Arle Lommel[**] and Hans Uszkoreit[*]

[*]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
firstname.lastname@dfki.de
[**]Common Sense Advisory (CSA Research), Massachusetts, USA
alommel@csa-research.com

Abstract

We present an alternative method of evaluating Quality Estimation systems, which is based on a linguistically-motivated Test Suite. We create a test-set consisting of 14 linguistic error categories and we gather for each of them a set of samples with both correct and erroneous translations. Then, we measure the performance of 5 Quality Estimation systems by checking their ability to distinguish between the correct and the erroneous translations. The detailed results are much more informative about the ability of each system. The fact that different Quality Estimation systems perform differently at various phenomena confirms the usefulness of the Test Suite.

1 Introduction

The evaluation of empirical Natural Language Processing (NLP) systems is a necessary task during research for new methods and ideas. The evaluation task is the last one to come after the development process and aims to indicate the overall performance of the newly built system and compare it against previous versions or other systems. Additionally, it also allows for conclusions related to the decisions taken for the development parameters and provides hints for improvement. Defining evaluation methods that satisfy the original development requirements is an ongoing field of research.

Automatic evaluation in sub-fields of Machine Translation (MT) has been mostly performed on given textual hypothesis sets, where the performance of the system is measured against gold-standard reference sets with one or more metrics (Bojar et al., 2017). Despite the extensive research on various automatic metrics and scoring methods, little attention has been paid to the actual content of the test-sets and how these can be adequate for judging the output from a linguistic perspective. The text of most test-sets so far has been drawn from various random sources and the only characteristic that is controlled and reported is the generic domain of the text.

In this paper we make an effort to demonstrate the value of using a linguistically-motivated controlled test-set (also known as a *Test Suite*) for evaluation instead of generic test-sets. We will focus on the sub-field of sentence-level Quality Estimation (QE) on MT and see how the evaluation of QE on a Test Suite can provide useful information concerning particular linguistic phenomena.

2 Related work

There have been few efforts to use a broadly-defined Test Suite for the evaluation of MT, the first of them being during the early steps of the technology (King and Falkedal, 1990). Although the topic has been recently revived (Isabelle et al., 2017; Burchardt et al., 2017), all relevant research so far applies only to the evaluation of MT output and not of QE predictions.

Similar to MT output, predictions of sentence-level QE have also been evaluated on test-sets consisting of randomly drawn texts and a single metric has been used to measure the performance over the entire text (e.g. Bojar et al., 2017). There has been criticism on the way the test-sets of the shared tasks have been formed with regards to the distribution of inputs (Anil and Fran, 2013), e.g. when they demonstrate a dataset shift (Quionero-Candela et al., 2009). Additionally, although there has been a lot of effort to infuse linguistically motivated features in QE (Felice and Specia, 2012), there has been no effort to evaluate their predictions from a linguistic perspective. To the best of our knowledge there has been no use of a Test Suite in order to evaluate sentence-level QE, or to inspect the predictions with regards to linguistic categories or specific error types.

3 Method

The evaluation of QE presented in this paper is based on these steps: (1) construction of the Test Suite with respect to linguistic categories; (2) selection of suitable Test Suite sentences; and (3) analysis of the Test Suite by existing QE systems and statistical evaluation of the predictions. These steps are analysed below, whereas a simplified example is given in Figure 1.

3.1 Construction of the Test Suite

The Test Suite has been developed by a professional linguist, supported by professional translators. First, the linguist gathers or creates error-specific paradigms (Figure 1, stage a), i.e. sentences whose translation has demonstrated or is suspected to demonstrate systematic errors by known MT engines. The aim is to have a representative amount of paradigms per error type and the paradigms are as short as possible in order to focus solely on one phenomenon under examination. The error types are defined based on linguistic categories inspired by the MQM error typology (Lommel et al., 2014) and extend the error types presented in Burchardt et al. (2017), with additional fine-grained analysis of sub-categories. The main categories for German-English can be seen in Table 2.

Second, the paradigms are given to several MT systems (Figure 1, stage b) to check whether they are able to translate them properly , with the aim to acquire a “pass” or a “fail” label accordingly. In an effort to accelerate the acquisition of these labels, we follow a semi-automatic annotation method using regular expressions. The regular expressions allow a faster automatic labelling that focuses on particular tokens expected to demonstrate the issue, unaffected from alternative sentence formulations. For each gathered source sentence the linguist specifies regular expressions (Figure 1, stage c) that focus on the particular issue: one positive regular expression that matches a successful translation and gives a “pass” label and an optional negative regular expression that matches an erroneous translation and gives a “fail” (for phenomena such as ambiguity and false friends). The regular expressions, developed and tested on the first translation outputs, are afterwards applied to all the alternative translation outputs (stage d) to acquire the automatic labels (stage e). Further modifications to the regular expressions were applied, if they did not properly match the new translation outputs. The automatically assigned labels were controlled in the end by a professional translator and native speaker of the target language (stage f). For the purposes of this analysis, we also assume that every sentence paradigm only demonstrates the error type that it has been chosen for and no other major errors occur.

MT type	proportion
neural	64.7%
phrase-based	26.8%
both (same output)	8.5%

Table 1: MT type for the translations participating in the final pairwise test-set

3.2 Selection of suitable Test Suite sentences

The next step is to transform the results so that they can be evaluated by existing sentence-level QE methods, since the Test Suite provides binary pass/fail values for the errors, whereas most sentence-level QE methods predict a continuous score. For this purpose, we transform the problem to a problem of predicting comparisons. We deconstruct the alternative translations of every source sentence into pairwise comparisons, and we only keep the pairs that contain one successful and one failing translation (Figure 1, stage g). Sentence-level QE systems will be given every pair of MT outputs and requested to predict a comparison, i.e. which of the two outputs is better (stage h). Finally, the QE systems are evaluated based on their capability to properly compare the erroneous with the correct outputs (stage i). The performance of the QE systems will be therefore expressed in terms of the accuracy over the pairwise choices.

4 Experiment

4.1 Data and systems

The current Test Suite contains about 5,500 source sentences and their rules with regular expressions for translating German to English. These rules have been applied for evaluating 10,800 unique

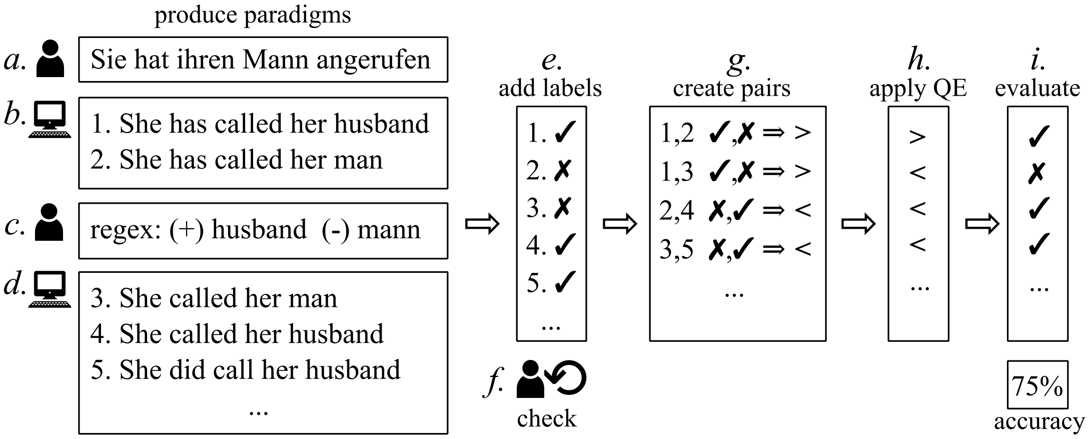

Figure 1: Example for the processing of test items for the lexical ambiguity of word "Mann"

MT outputs (MT outputs with the exact same text have been merged together). These outputs have been produced by three online commercial systems (2 state-of-the-art neural MT systems and one phrase-based), plus the open-source neural system by Sennrich et al. (2017). After creating pairs of alternative MT outputs that have a different label (Section 3.2) the final test-set contains 3,230 pairwise comparisons based on the translations of 1,582 source sentences. The MT types of the translations participating in the final test-set can be seen in Table 1.

For this comparative study we evaluate existing QE systems that were freely available to train and use. In particular we evaluate the baseline the following 6 systems:

- **B17:** The baseline of the shared task on sentence-level QE (Bojar et al., 2017) based on 17 black-box features and trained with Support Vector Regression (SVR) to predict continuous HTER values

- **B13:** the winning system of the shared task on QE ranking (Bojar et al., 2013; Avramidis and Popović, 2013) based on 10 features, trained with Logistic Regression with Stepwise Feature Selection in order to perform ranking. Despite being old, this system was chosen as it is the latest paradigm of Comparative QE that has been extensively compared with competitive methods in a shared task

- **A17:** three variations of the state-of-the-art research on Comparative QE (Avramidis, 2017), all three trained with a Gradient Boosting classifier. The *basic* system has the same feature set as B13, the *full* system contains a wide variety of 139 features and the *RFECV* contains the 25 highest ranked features from the full feature set, after running Recursive Feature Elimination with an SVR kernel.

The implementation was based on the open-source tools Quest (Shah et al., 2013) and Qualitative (Avramidis, 2016).

4.2 Results

Here we present the evaluation of the QE systems when applied on the Test Suite. The accuracy achieved by each of the 6 QE systems for the 14 error categories can be seen in Table 2.

First, it can be noted that the **quantity of evaluated samples** varies a lot and, although the original aim was to have about 100 samples per category, most of the neural outputs succeeded in the translations of the issues and therefore were not included in the test-set with the "pass/fail" comparisons. Obviously, conclusions for those error categories with few samples cannot be guaranteed.

Second, one can see that the **average scores** range between 52.1% and 57.5% (achieved by B13) which are nevertheless relatively low. This may be explained by the fact that all QE systems have been developed in the previous years with the focus on "real text" test-sets. The Test Suite on the contrary is not representative of a real scenario and has a different distribution than the one expected from real data. Additionally, many of the linguistic phenomena of the Test Suite may have few or no occurrences on the development data of the QE systems. Finally, all QE systems have been devel-

		B17	B13		A17	
	amount	baseline	winning	basic	RFECV	full
Ambiguity	89	58.4	64.0	**73.0**	69.7	62.9
Composition	75	58.7	77.3	**80.0**	72.0	77.3
Coordination & ellipsis	78	53.8	**73.1**	71.8	71.8	70.5
False friends	52	38.5	32.7	**48.1**	38.5	42.3
Function word	126	33.3	**38.9**	35.7	32.5	34.9
Long distance dep. & interrogatives	266	52.3	63.9	60.2	63.9	**65.8**
Multi-word expressions	43	32.6	**44.2**	32.6	39.5	39.5
Named entity & terminology	55	50.9	54.5	56.4	58.2	**60.0**
Negation	13	38.5	53.8	**76.9**	**76.9**	**76.9**
Non-verbal agreement	45	40.0	**57.8**	53.3	**57.8**	53.3
Punctuation	138	11.6	29.7	**32.6**	28.3	27.5
Subordination	46	41.3	43.5	**47.8**	45.7	**47.8**
Verb tense/aspect/mood/type	2137	56.6	**59.4**	55.5	57.3	57.7
Verb valency	67	50.7	55.2	50.7	58.2	**62.7**
Total	3230	52.1	**57.5**	55.0	56.1	56.7
weighed		44.1	53.4	55.3	55.0	**55.6**

Table 2: QE accuracy (%) per error category

oped in the previous years with the focus on rule-based or phrase-based statistical MT and therefore their performance on MT output primarily from neural systems is unpredictable.

We also report scores averaged not out of the total amount of the samples, but instead giving equal importance to each error category. These scores indicate a different winner: the full system of A17. However, due to the distributional shift of the Test Suite, there is limited value in drawing conclusions from average scores, since the aim of the Test Suite is to provide a qualitative overview of the particular linguistic phenomena.

When it comes to **particular error categories**, the three systems B13, A17-basic and A17-full seem to be complementary, achieving the highest score for 5 different error categories each. The systems B17 and A17-RFECV lack a lot in their performance. The highest category score is achieved for the phenomenon of *Composition* (compounds and phrasal verbs) by A17-basic, followed by *negation* (albeit with very few samples) at 76.9%. A17-basic is also very strong in *ambiguity*, achieving 73%. The 4 systems B13 and A17 perform much better concerning *long-distance relationships*, which may be attributed to the parsing and grammatical features they contain, as opposed to the B17 which does not include parsing. Finally, A17-full does better with *named entities* and *terminology*, possibly because its features include alignment scores from IBM model 1.

We notice that **verb tenses, aspects, moods and types** comprise a major error category which contains more than 2,000 samples. This enables us to look into the subcategories related to the verbs. The performance of the systems for different tenses can be seen in Table 3, where B17 and B13 are the winning systems for 5 categories each. The tense with the best performance is the *future II subjunctive II* with a 78% accuracy by B13. Despite its success in the broad spectrum of error categories, A17-full performs relatively poorly on verb tenses.

Finally, Table 4 contains the accuracy scores for **verb types**. A17-full does much better on verb types, with the exception of the *negated modal* which gets a surprising 70.3% accuracy from B17.

5 Conclusion and further work

In this paper we demonstrated the possibility of performing evaluation of QE by testing its predictions on a fine-grained error typology from a Test Suite. In this way, rather than judging QE systems based on a single score, we were able to see how each QE system performs with respect to particular error categories. The results indicate that no system is a clear winner, with three out of the 5 QE systems to have complementary results for

		B17	B13	⌈	A17	⌉
	amount	baseline	winning	basic	RFECV	full
future I	297	**58.9**	**58.9**	52.5	50.5	51.5
future I subjunctive II	249	**62.7**	52.6	45.0	51.4	53.0
future II	158	39.2	56.3	**60.1**	58.2	53.2
future II subjunctive II	168	32.7	**78.0**	74.4	68.5	75.6
perfect	294	55.4	**56.8**	49.3	55.8	54.8
pluperfect	282	**72.7**	65.6	64.9	69.9	68.1
pluperfect subjunctive II	159	52.2	53.5	**55.3**	52.8	**55.3**
present	286	**58.0**	54.9	51.4	51.0	52.8
preterite	105	61.0	**68.6**	53.3	67.6	**68.6**
preterite subjunctive II	88	**62.5**	61.4	58.0	53.4	55.7

Table 3: QE accuracy (%) on error types related to verb tenses

		B17	B13	⌈	A17	⌉
	amount	baseline	winning	basic	RFECV	full
Ditransitive	275	46.9	57.8	55.6	56.4	**60.0**
Intransitive	171	42.1	**69.6**	57.3	59.1	64.3
Modal	473	63.4	67.2	57.9	66.6	**67.2**
Modal negated	657	**70.3**	49.9	47.2	46.0	46.3
Reflexive	376	44.7	61.2	61.2	**62.2**	58.5
Transitive	134	39.6	68.7	69.4	64.9	**68.7**

Table 4: QE accuracy (%) on error types related to verb types

all the error categories. The fact that different QE systems with similar overall scores perform differently at various phenomena confirms the usefulness of the Test Suite for understanding their comparative performance.

Such linguistically-motivated evaluation can be useful in many aspects. The development or improvement of QE systems may use the results about the found errors in order to introduce new related features. The development may also be aided by testing these improvements on an isolated development set.

Further work should include the expansion of the Test Suite with more samples in the less-populated categories and support for other language pairs. Finally, we would ideally like to broaden the comparison among QE systems, by including other state-of-the-art ones that unfortunately were not freely available to test.

Acknowledgments

Part of this work has received funding from the EU Horizon 2020 research and innovation program QT21 under grant agreement N^o 645452.

References

Guillaume Wisniewski Anil and Kumar Singh Fran. 2013. Quality Estimation for Machine Translation: Some Lessons Learned. *Machine Translation* 27(3-4):213–238. https://link.springer.com/article/10.1007/s10590-013-9141-9.

Eleftherios Avramidis. 2016. Qualitative: Python Tool for MT Quality Estimation Supporting Server Mode and Hybrid MT. *The Prague Bulletin of Mathematical Linguistics (PBML)* 106:147–158. https://ufal.mff.cuni.cz/pbml/106/art-avramidis.pdf.

Eleftherios Avramidis. 2017. Comparative Quality Estimation for Machine Translation: Observations on machine learning and features. *Proceedings of the 20th Annual Conference of the European Association for Machine Translation, The Prague Bulletin of Mathematical Linguistics* (108):307–318. https://doi.org/10.1515/pralin-2017-0029.

Eleftherios Avramidis and Maja Popović. 2013. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 329–336. http://www.aclweb.org/anthology/W13-2240.

Ondej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 12–58. https://doi.org/10.3115/1626431.1626433.

Ondej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 169–214. http://www.aclweb.org/anthology/W17-4717.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics* 108:159–170. https://doi.org/10.1515/pralin-2017-0017.

Mariano Felice and Lucia Specia. 2012. Linguistic Features for Quality Estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada, pages 96–103. http://www.aclweb.org/anthology/W12-3110.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1704.07431.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. *Proceedings of the 13th conference on Computational linguistics -* 2:211–216. https://doi.org/10.3115/997939.997976.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT-14)*. Croatian Language Technologies Society, European Association for Machine Translation, pages 165–172. http://www.mt-archive.info/10/EAMT-2014-Lommel.pdf.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil Lawrence. 2009. *Dataset shift in machine learning*. MIT Press.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 389–399. http://www.aclweb.org/anthology/W17-4739.

Kashif Shah, Eleftherios Avramidis, Ergun Biçici, and Lucia Specia. 2013. QuEst: Design, Implementation and Extensions of a Framework for Machine Translation Quality Estimation. *The Prague Bulletin of Mathematical Linguistics* 100:19–30. https://doi.org/10.2478/pralin-2013-0008.PBML.