EVALD Reference-Less Discourse Evaluation for WMT18

Ondřej Bojar **Jiří Mírovský** **Kateřina Rysová** **Magdaléna Rysová**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
<surname>@ufal.mff.cuni.cz

Abstract

We present the results of automatic evaluation of discourse in machine translation (MT) outputs using the EVALD tool. EVALD was originally designed and trained to assess the quality of *human* writing, for native speakers and foreign-language learners. MT has seen a tremendous leap in translation quality at the level of sentences and it is thus interesting to see if the human-level evaluation is becoming relevant.

1 Introduction

The output quality of machine translation has substantially improved in the last few years thanks to the neural models (NMT). In some setups, NMT systems may even surpass the quality of human reference translations *if evaluated at the level of individual sentences*. The natural next step is (1) to start evaluating MT using larger pieces of texts, e.g. whole documents, and (2) to evaluate using methods suitable for the text quality produced by humans.

Our contribution to the WMT18 test suites responds to both of these goals. We experiment with the application of automatic, reference-less evaluation of text quality which was originally designed to evaluate texts written by humans. In this exploratory study, we do not have the human resources for a contrastive manual evaluation of the texts. We thus limit the comparison to overall MT system quality as provided by WMT.

In Section 2, we briefly describe the tool we use, EVALD. Section 3 describes the texts and MT system used. Section 4 provides and discusses the empirical results and we conclude in Section 5.

2 Evaluating Discourse

EVALD (Evaluator of Discourse)[1] was used for the automatic evaluation of the translated texts. There are two main versions of EVALD: EVALD for native speakers of Czech ("L1") and EVALD for non-native speakers ("L2"). The versions share the same features but differ in training texts.

EVALD L1 was trained on 1118 essays written by native speakers, while EVALD L2 was trained on 945 essays written by learners of Czech as a foreign language. Both systems use the same 180 features that can be divided into two types: (i) shallow features that use information from lower layers of language description, namely spelling, vocabulary, morphology and syntax, and (ii) deep text features directly related to surface coherence and reaching also beyond the sentence boundaries, namely coreference, discourse connectives diversity, discourse connectives quantity, and sentence information structure. Details about the systems can be found in Novák et al. (2018), Rysová et al. (2018), Novák et al. (2017), or Rysová et al. (2017).

We expect EVALD L2 to work better because it was designed and trained for evaluation of texts that are usually not fully coherent. The same aspect is expected by the automatically translated texts – they can be sometimes disrupted from the linguistic point of view.

EVALD L1 and L2 also differ in the class labels assigned. We normalize both of them to assign scores from 1 (worst) to 5 or 6 (best; L1 uses 5 classes, L2 uses 6 classes).

3 Data

Since the domain of WMT18 Shared Translation Task is news, we needed to find a different input

[1] https://lindat.mff.cuni.cz/services/evald-foreign/

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 545–549

https://doi.org/10.18653/v1/W18-6432

	ENG	NRE	PSY	SOC	EDU	HIS	IOE	PHI	POL	CLS	ECO	LIN	NUR	BIO	CEE	MEC	PHY
Creative Writing	3/1	1/-	1/-	1/-													
Critique	4/-		2/2	-/2	2/-	-/1	1/1	2/-	3/-								
Essay	4/1		-/1	-/2	2/-	1/1		-/2	-/1	3/-	1/-	1/-	1/-				
Proposal		1/-	1/-	-/2	3/-		1/1		2/-			2/1	1/2	-/1			
Report		-/2	-/2	-/6	2/3		-/3		-/2			-/3	-/1	1/-	2/-	-/1	-/1
Research Paper			2/-	4/-					1/-		-/1			-/2	-/2		

Table 1: Texts in our test suite by genre and domain. The numbers indicate texts written by a native/non-native English speaker.

texts which matches more closely to domain that EVALD is trained for.

3.1 Evaluated Texts

We selected Michigan Corpus of Upper-Level Student Papers (MICUSP),[2] an open-source collection of original English texts developed at the University of Michigan (English Language Institute). MICUSP contains about 830 papers (2.6 million words). The texts come from four academic areas: Humanities and Arts, Social Sciences, Biological and Health Sciences, Physical Sciences. At the same time, various text genres are present (argumentative essay, creative writing, critique/evaluation, proposal report, research paper, response paper). Authors of the papers are final year undergraduate and graduate students who reached an A grade. The corpus contains texts written by the native as well as non-native speakers of English. The overview of the MICUSP texts selected for evaluation is presented in Table 1.

The genre that should fit EVALD best is creative writing. We thus specifically extracted all 7 texts labelled as creative writing. To further extend our test suite, we selected texts of suitable length across the genres and domains, as summarized in Table 1. In total, there are 56 texts written by native speakers and 51 texts written by non-native speakers.

We segmented the texts into individual sentences and manually edited them to correct any errors in segmentation, to remove auxiliary segments like "[Figure]" and to abbreviate them occasionally by removing e.g. inline tables.

3.2 MT Systems Used

The final texts were included in inputs of MT systems participating in the WMT18 News Translation Task. In addition to the "primary" systems CUNI Transformer, UEDIN and the online systems, we also added three baseline (contrastive) systems: CUNI Chimera, CUNI Chimera noDepfix and CUNI Moses.

CUNI Moses is a phrase-based MT system (Koehn et al., 2007) trained on very large data and domain-adapted for the news text. CUNI Chimera (Bojar et al., 2013) is a hybrid MT system combining the outputs of transfer-based TectoMT (Žabokrtský et al., 2008) and recently also neural MT outputs from Nematus (Sennrich et al., 2017) and Neural Monkey (Helcl et al., 2018). The backbone of Chimera is nevertheless phrase-based, so Chimera suffers from the standard problems of fluency. Depfix (Rosa et al., 2012) is a rule-based grammar correction system that served very well as the last step of Chimera prior to NMT. For a contrast, we also provide the outputs of Chimera without this rule-based component.

CUNI Transformer (Popel and Bojar, 2018) is a highly optimized NMT system based on the non-recurrent architecture of Transformer (Vaswani et al., 2017). Based on the preliminary evaluation, CUNI Transformer is expected to perform comparably or better than humans when evaluating individual sentences in isolation.

UEDIN is a 4-way ensemble of deep RNN system, running left-to-right and reranked with 4 deep right-to-left systems. It uses subword units (BPE) and back-translation. The other systems are commercial ones and their description is not available.

The manual evaluation of WMT18 is still in progress, so what we can provide now are only automatic scores as reported in `matrix.statmt.org`, see Table 2. None of the WMT18 evaluations will be strictly comparable to ours due to the difference in the domain and the set of sentences. Nevertheless, it is still the best indication of MT output quality we can get.

4 Evaluation

We apply EVALD to all the MT outputs and also to the source. No Czech reference is available for the texts, so we take the source as the lower bound:

[2] http://micusp.elicorpora.info/

System	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
CUNI Transformer	26.6	26.0	0.638	0.567	0.532
UEDIN	24.0	23.4	0.666	0.554	0.550
CUNI Chimera noDepFix	21.0	19.8	0.703	0.528	0.600
CUNI Chimera	20.8	19.2	0.704	0.522	0.605
CUNI Moses	17.5	16.4	0.739	0.509	0.632

Table 2: Automatic results of WMT18 English-Czech systems as listed at http://matrix.statmt.org/matrix/systems_list/1883.

EVALD version	L1	L2
CUNI Transformer	5.00±0.00	5.02±0.91
CUNI Chimera noDepFix	5.00±0.00	4.92±0.88
UEDIN	5.00±0.00	4.77±0.89
online-B	5.00±0.00	4.76±0.87
CUNI Moses	4.97±0.29	4.69±0.83
online-A	5.00±0.00	4.60±0.81
CUNI Chimera	5.00±0.00	4.58±0.80
online-G	4.97±0.29	4.58±0.81
Source	1.00±0.00	1.00±0.00

Table 3: Overall EVALD scores for individual MT systems. L1: EVALD for native speakers with 5 being the best mark, L2: EVALD for non-natives with 6 being the best possible mark.

	EVALD L2 Score	# Docs	#
Creative Writing	6.00±0.00	7	56
Report	4.72±0.84	29	289
Essay	4.67±0.89	21	153
Critique	4.65±0.90	20	136
Research Paper	4.59±0.70	12	90
Proposal	4.52±0.66	18	132

Table 4: Results for individual genres.

	EVALD L2 Score	#	# Docs
HIS	5.48±0.89	27	3
ENG	5.23±0.97	83	13
NRE	5.11±0.63	35	4
IOE	5.03±0.79	65	7
PSY	4.83±0.97	88	11
SOC	4.79±0.91	160	17
BIO	4.74±0.60	38	4
CEE	4.62±0.61	32	4
ECO	4.56±0.73	16	2
EDU	4.55±0.86	78	12
POL	4.48±0.80	63	9
LIN	4.44±0.73	59	7
NUR	4.37±0.49	43	5
MEC	4.36±0.50	11	1
PHY	4.36±0.50	11	1
CLS	4.27±0.46	15	3
PHI	4.00±0.00	32	4

Table 5: Results for individual domains.

	EVALD L2 Score	#	# Docs
Native Speaker	4.86±0.93	298	56
Non-Native Speaker	4.68±0.82	558	51

Table 6: Results depending on whether the author of the English original was an English native speaker.

EVALD, trained for Czech, should very much dislike the original English text.

The overall EVALD score across the 107 texts produced by each MT system is listed in Table 3. Clearly, the L1 version of EVALD aimed at native speakers is non-discerning. All systems get almost the same score. It is actually the best possible score, but this tells us primarily that the system trained for L1 is not suitable for our setting. Only the source gets the worst possible score.

The L2 version is more interesting. As expected, English Source receives the worst rating, 1.0 with no variance at all. MT systems score around 4 or 5. While this is a clear overestimation of the text quality (6 would be the best score and e.g. phrase-based MT Moses gets 4.69), it reveals some differences between the systems.

We thus explore only EVALD L2 in the following.

Table 4 lists EVALD L2 scores for individual genres across MT systems; Source was not considered. The columns "#" and "# Docs" specify the size of the sample in terms of individual scorings and distinct documents, respectively.

We see that all 56 translations of the 7 documents of Creative Writing seemed excellent. Again, EVALD is non-discerning in this setting. Other genres exhibit some divergence in scores. Since all the genres differ from the news texts that the MT systems are geared towards, it is not easy to explain the stability of the score in Creative Writing. Possibly, EVALD is checking many shallow discourse features (e.g. the presence of a certain variety of conjunctions) and our texts in Creative Writing superficially include the required diversity, and this diversity is preserved by all MT systems.

Table 5 looks at text domains. There is a reasonable variance across the translations and texts (except PHIlosophy) but it is again difficult to come up with a unified view. For instance, natural sciences like BIOlogy or PHYsics span a wide range

	Discourse-Specific	Other	All
CUNI Transformer	4.56±1.18	4.79±1.16	5.02±0.91
CUNI Chimera noDepFix	4.52±1.15	≀ 4.86±1.17	4.92±0.88
UEDIN	4.52±1.15	4.86±1.09	4.77±0.89
online-B	4.39±1.17	4.82±1.12	4.76±0.87
online-G	4.35±1.15	4.68±1.21	4.58±0.81
online-A	4.34±1.12	4.66±1.28	≀ 4.60±0.81
CUNI Moses	4.30±1.24	≀ 4.69±1.28	≀ 4.69±0.83
CUNI Chimera	3.98±1.36	4.66±1.20	4.58±0.80
Source	1.86±1.65	2.00±1.73	1.00±0.00

Table 7: Comparison of EVALD L2 scores using discourse-specific (deep) features, other (shallow) features, and all features. Vertical tildes mark differences in rank in comparison with the rank given by the discourse-specific features.

	Avg. var. of scores
across nativeness	0.88
across MT systems	0.85
across genre	0.67
across domain	0.67

Table 8: Variance in EVALD L2 scores across various aspects of our test suite.

of ranks, as humanities do (HIStory or the mentioned PHIlosophy).

Table 6 documents the effect of the mother tongue of the author of the original English text before the translation.

Table 7 compares EVALD L2 scores in three experimental settings: using only the deep text features (marked discourse-specific in the table), shallow features (marked other) and all features.[3] Vertical tildes mark differences in rank in comparison with the rank given by the deep text features. Agreement in five first ranks using the deep features and all features indicates that the full version of EVALD (i.e. using all features) really evaluates the translation systems based on the quality of the text coherence, rather than on the basis of shallow features.

Table 8 summarizes the variance of EVALD scores according to individual aspects captured in the previously mentioned tables. The highest variance of the scores appeared in the aspect of nativeness of the text author.

The second most diverse results are across MT systems. The evaluation proposed here thus seems as a promising research direction, although a careful analysis of EVALD features and their adaptation will be needed to obtain more discerning evaluation. Finally, the genre and domain of the original text also play a role but this is always to be expected.

[3] See Section 2 for the list of features.

5 Conclusion

We presented the results of automatic evaluation of Czech text quality applied to the output of generally good MT systems translating from English into Czech.

The results indicate that EVALD, as now trained for human-authored texts, is ineffective in its version for native speakers. However, EVALD version for non-natives has a rather promising potential for evaluating automatic translations because it allows distinguishing individual MT systems.

The most diversity of scores can be attributed to the nativeness of the author of the original text. We conclude that the examined MT systems in general preserve sufficient traits of source text quality for this.

EVALD-style of evaluation seems promising because the second most differentiating aspect is the MT system used. Further exploration of EVALD features as well as a direct comparison with manual assessment of translation quality are, however, necessary to make EVALD a useful MT evaluation method.

Acknowledgments

This work has been supported by the grants 18-24210S, GA17-06123S and GA17-03461S of the Czech Science Foundation, SVV 260 453, and using language resources and tools distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

References

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on*

Statistical Machine Translation, pages 92–98, Sofia, Bulgaria. Association for Computational Linguistics.

Jindřich Helcl, Jindřich Libovický, Tom Kocmi, Tomáš Musil, Ondřej Cífka, Dušan Variš, and Ondřej Bojar. 2018. Neural monkey: The current state and beyond. In *The 13th Conference of The Association for Machine Translation in the Americas, Vol. 1: MT Researchers' Track*, pages 168–176, Stroudsburg, PA, USA. The Association for Machine Translation in the Americas, The Association for Machine Translation in the Americas.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Michal Novák, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018. Topic-focus articulation: A third pillar of automatic evaluation of text coherence. conference paper. In *Proceedings of MICAI 2018*. Springer LNAI – Lecture Notes in Artificial Intelligence, in press.

Michal Novák, Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2017. Incorporating coreference to automatic evaluation of coherence in essays. In *Statistical Language and Speech Processing*, number 10583 in Lecture Notes in Computer Science, pages 58–69, Cham, Switzerland. Claude Chappe Informatics Institute at University of Le Mans, Springer International Publishing.

Martin Popel and Ondej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Rudolf Rosa, David Mareček, and Ondej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada. Association for Computational Linguistics.

Kateřina Rysová, Magdaléna Rysová, Jiří Mírovský, and Michal Novák. 2017. Introducing EVALD software applications for automatic evaluation of discourse in czech. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 634–641, Šumen, Bulgaria. Bulgarian Academy of Sciences, INCOMA Ltd.

Magdaléna Rysová, Kateřina Rysová, Jiří Mírovský, and Michal Novák. 2018. Practicing students writing skills through elearning: Automated evaluation of text coherence in czech. In *EDULEARN18 Proceedings*, pages 1963–1970, Valencia, Spain. IATED Academy, IATED Academy.

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*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA.