Using MT-ComparEval

Roman Sudarikov,[α] **Martin Popel,**[α] **Ondřej Bojar,**[α] **Aljoscha Burchardt,**[β] **Ondřej Klejch**[α,γ]
[α] Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[β] German Research Center for Artificial Intelligence (DFKI), Language Technology Lab
[γ] Centre for Speech Technology Research, University of Edinburgh
{sudarikov,popel,bojar,klejch}@ufal.mff.cuni.cz, aljoscha.burchardt@dfki.de

Abstract

The paper showcases the MT-ComparEval tool for qualitative evaluation of machine translation (MT). MT-ComparEval is an open-source tool that has been designed in order to help MT developers by providing a graphical user interface that allows the comparison and evaluation of different MT engines/experiments and settings. The tool implements several measures that represent the current best practice of automatic evaluation. It also provides guidance in the targeted inspection of examples that show a certain behavior in terms of n-gram similarity/dissimilarity with alternative translations or the reference translation. In this paper, we provide an applied, "hands-on" perspective on the actual usage of MT-ComparEval. In a case study, we use it to compare and analyze several systems submitted to the WMT 2015 shared task.

Keywords: Machine Translation Evaluation, Analysis of MT Output

1. Introduction

The MT *development* cycle is well supported by several sophisticated pipeline tools such as the Experiment Management System EMS (Koehn, 2010) that comes with the Moses toolkit. While these pipelines support intermediate evaluation steps, there has been a lack of versatile tool support for detailed, qualitative evaluation steps, in particular for:

- Systematically documenting *significant quantitative* changes in terms of various automatic measures over a potentially large number of different system types, combinations, and variants, and
- *Qualitatively* analyzing the effects of changes in the systems integrated with the above.

The evaluation interface of EMS shows some similarity with the tool to be described in this paper, but it is tightly connected with the training pipeline and optimized for the Moses statistical machine translation (SMT) scripts. MT-ComparEval in contrast provides more flexibility, since the evaluation interface can be run independently of the production of translation systems.[1]
The automatic quantitative evaluation of MT is supported by different metrics such as BLEU, Meteor, TER (Agarwal and Lavie, 2008). A separate issue with these metrics is that they are usually implemented in unrelated collections of scripts which, in addition to unnecessary burden of tool installation, easily leads to difficulties with replicability: we can get different outcomes for the same metric from different implementations. The bigger problem, however, is still the question of how to perform qualitative evaluation.
MT-ComparEval addresses the problems described above. Klejch et al. (2015) have introduced the tool from a general perspective, focusing on how to deal with automatic measures. In this paper, we will present it from the "hands-on" view of a researcher who compares several machine translation systems with the goal of getting deeper insights into these systems than "System A is better than system B by 1.5 BLEU score" and possibly also with the goal to improve some of the systems.

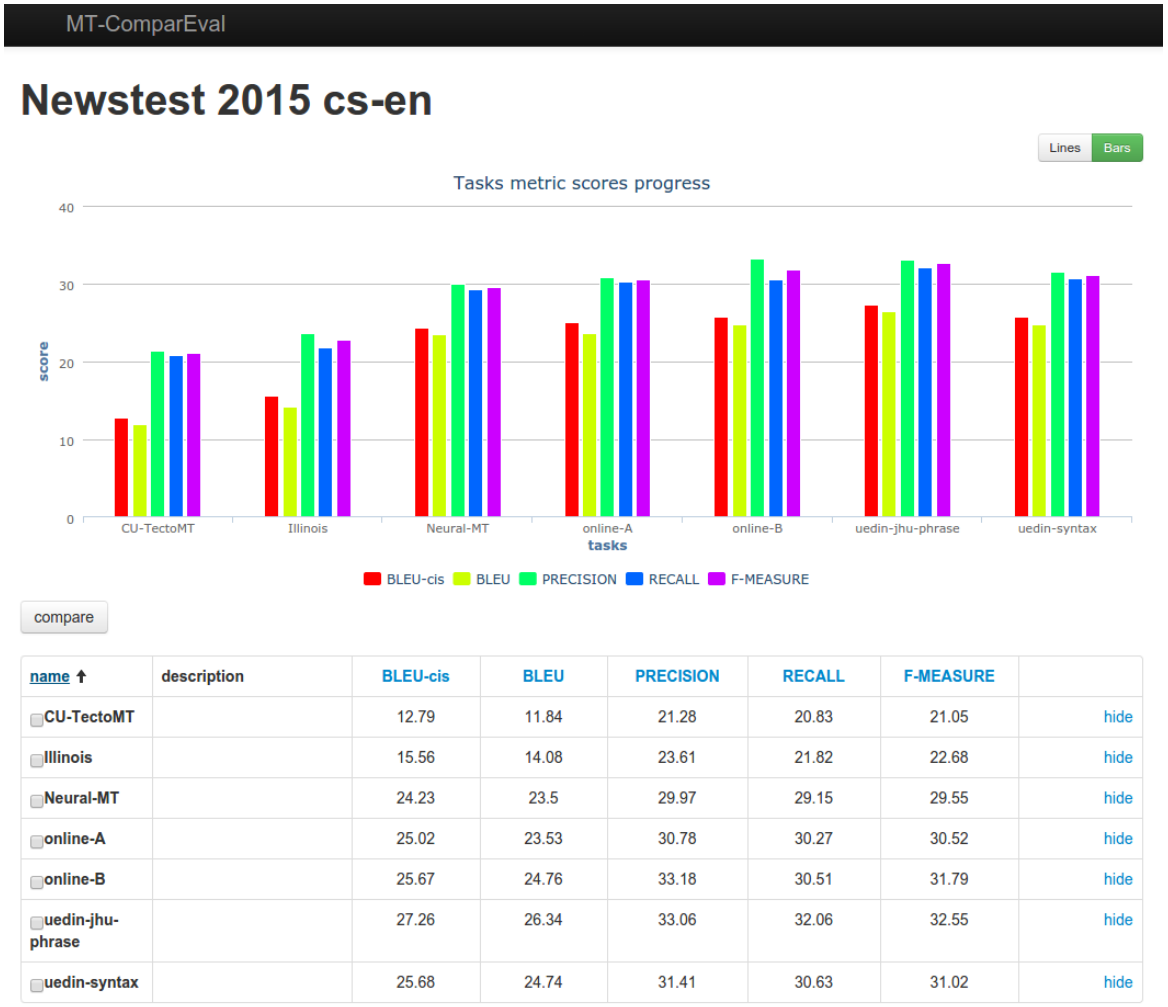
[1] See Klejch et al. (2015) for a more detailed discussion.

2. MT-ComparEval

MT-ComparEval, the open-source tool described in this article, has been designed in order to help MT developers by providing a graphical user interface that allows the comparison and evaluation of different MT engines/experiments and settings through the use of several measures that represent the current best practice. The user interface is web-based and backed by a server side of the tool.
This paper won't dwell on the internal structure of the tool, but rather point out certain main features which will later be used in qualitative evaluation of machine translation systems. These features include:

- Integration of different evaluation metrics – by default MT-ComparEval configuration includes precision, recall and F-measure (all based on arithmetic average of 1-grams up to 4-grams), BLEU score and Brevity penalty (Papineni et al., 2002). It can produce also Hjerson (Popović, 2011) evaluation scores "out-of-the-box" when enabled in the configuration file.
- Focus on pairwise comparisons of MT systems – so strengths and weaknesses of one system are shown relative to another system.[2]

[2] We consider the pairwise comparisons a great advantage of MT-ComparEval (compared to other tools) because it focuses on the errors that are more likely to be repairable (because the second MT system was able to translate these correctly), instead of simply focusing on n-grams/sentences that are generally difficult to translate. If only one system is available, it is still possible to analyze it with MT-ComparEval by selecting the reference translation as the second "system". We plan to promote this feature in the interface because it may be useful *per se* (even if more systems are available).



name	description	BLEU-cis	BLEU	PRECISION	RECALL	F-MEASURE	
CU-TectoMT		12.79	11.84	21.28	20.83	21.05	hide
Illinois		15.56	14.08	23.61	21.82	22.68	hide
Neural-MT		24.23	23.5	29.97	29.15	29.55	hide
online-A		25.02	23.53	30.78	30.27	30.52	hide
online-B		25.67	24.76	33.18	30.51	31.79	hide
uedin-jhu-phrase		27.26	26.34	33.06	32.06	32.55	hide
uedin-syntax		25.68	24.74	31.41	30.63	31.02	hide

Figure 1: "Experiment" screen with overview of all tasks.

- Bootstrap resampling – MT-ComparEval automatically generates bootstrap samples and computes the p-value for systems comparison and confidence intervals for all the produced evaluation scores.
- Confirmed and unconfirmed n-grams – the tool presents top 10 n-grams (for n=1,2,3,4) where the two systems differ with respect to correctness of translation (as measured by the reference translation), that is n-grams that are responsible for the difference in BLEU scores. Full explanation is given in Section 3.4.
- Sentence comparison – MT-ComparEval provides a graphically rich interface for sentence by sentence comparison of systems' outputs.
- Accessibility – this tool can be easily installed and run locally (see "Installation" section at `https://github.com/choko/MT-ComparEval`).

In further sections, we will use MT-ComparEval terms "Experiment" and "Task" to refer to whole comparison and each system's output, respectively.

3. Using MT-ComparEval Step by Step

As a running example for the rest of the paper, we use a set of systems from the WMT2015 shared task (Bojar et al., 2015), Czech→English translation task.[3] All the described observations were done using the public MT-ComparEval server with WMT translations `http://wmt.ufal.cz`.[4] We encourage the readers to navigate to the "Newstest 2015 cs-en" experiment and try all the described steps.

3.1. Experiment Screen

Figure 1 shows the main screen of an "experiment", which lists the results of all "tasks" (MT systems' outputs) in this experiment. Figure 2 shows the same screen, where we

[3] `http://www.statmt.org/wmt15/translation-task.html`

[4] The buttons for uploading and deleting experiments and tasks are disabled at `http://wmt.ufal.cz`. Local installation of MT-ComparEval can be configured to show these buttons or to permanently monitor a `data` directory for new experiments and tasks, which is suitable for integrating MT-ComparEval into an MT development pipeline.

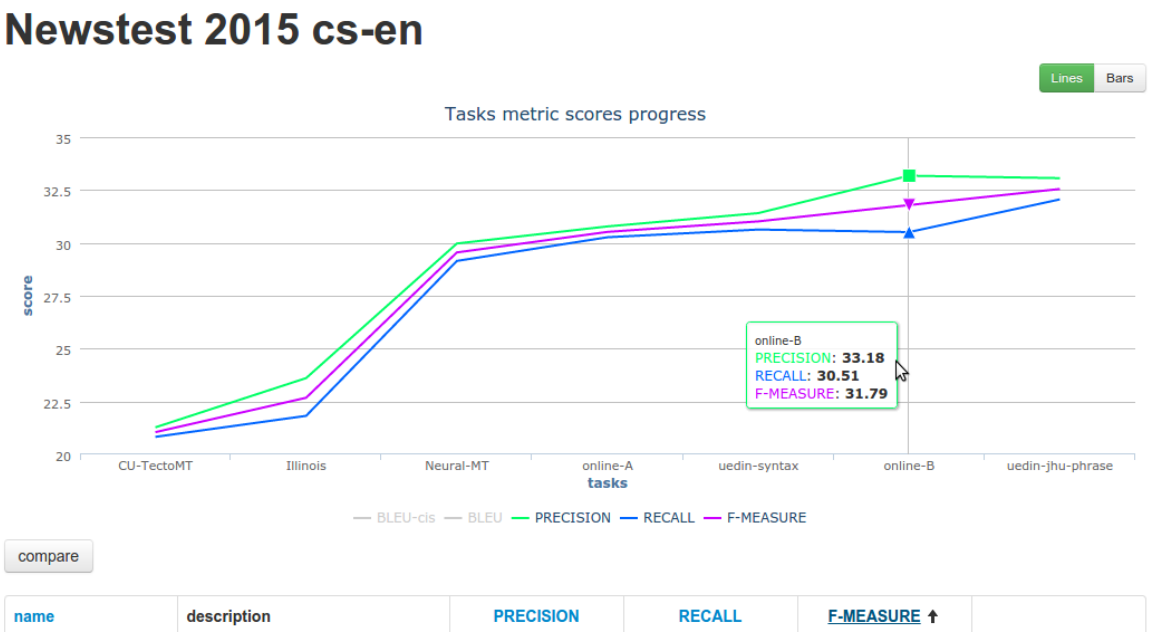


Figure 2: "Experiment" screen with Precision, Recall and F-measure (which was used for sorting the tasks).

have a) clicked on "F-MEASURE" in the table header, so the tasks are sorted according to this metric, b) clicked on the two variants of BLUE score under the graph to hide these metrics, so only Precision, Recall and F-measure are shown (and the graph y-axis is rescaled), and c) switched the graph type from bars to lines, so we can better see the differences between the metrics and check e.g. if some line segments are sloped down, which means disagreement with the F-measure used for sorting.

Metrics disagreement This actually happens for the top three systems, where Online-B is the best system according to Precision, second according to F-measure and third according to Recall.

Precision and Recall mismatch This is related to the fact that Online-B has Precision notably higher than Recall, while other systems have the difference much smaller. This may indicate that Online-B produces shorter translations and prefers to skip parts where the translation is not certain. This hypothesis can be checked by selecting Online-B and UEdin-jhu-phrase for pairwise comparison (see Section 3.2.) and looking at the sentences sorted according to RECALL or BREVITY-PENALTY (or even the default BLEU). See Figure 3 with two sentences where phrases *couldn't* and *high schools* were omitted in the Online-B translation.

Casing problems Figure 1 also shows that case-insensitive BLEU (BLEU-cis) is slightly lower than case-sensitive BLEU for all systems. The biggest difference is for Online-A (almost 1.5 BLEU points).[5] This indicates a problem with upper-casing.

3.2. Sentences Pane

MT-ComparEval focuses on comparing two tasks (systemA and systemB). After marking the tasks's checkboxes in the Experiment screen and clicking "Compare", a screen with four panes is shown: Sentences, Statistics, Confirmed n-grams and Unconfirmed n-grams, which are described in the following subsections.

The Sentences pane (Figure 3) shows all sentences from the given testset sorted according to the differences in the chosen sentence-level metric scores. This means that the sentences shown at the top are those where systemB outperforms systemA the most.[6] Such a view is very useful when checking for regressions of new versions of an MT system against a baseline or a previous version of the same system, but it is useful also when comparing different systems.

Color highlighting A set of checkboxes allow to highlight differences between the two systems in several ways:

- **Confirmed n-grams** are n-grams occurring both in the system output and in the reference.[7] These are marked with light yellow (Online-B) and blue (UEdin-jhu-phrase) background. The confirmed n-grams are highlighted also in the reference, where light green color marks n-grams occurring in both system (e.g. "Why" in the first sentence in Figure 3).

[5]This could be better seen when switching to the line graph and showing only BLEU and BLEU-cis. Online-A has one of the biggest differences in these two metrics also in other translation directions in WMT15: de-en, hi-en, fr-en and ru-en.

[6]The metric used for sorting and the increasing/decreasing ordering can be changed in the upper right corner.

[7]If a given n-gram occurs e.g. three times in the system output and only twice in the reference, a heuristic algorithm (based on the longest common subsequence) is used to select two occurrences of the n-gram that will be marked as confirmed in the system output.

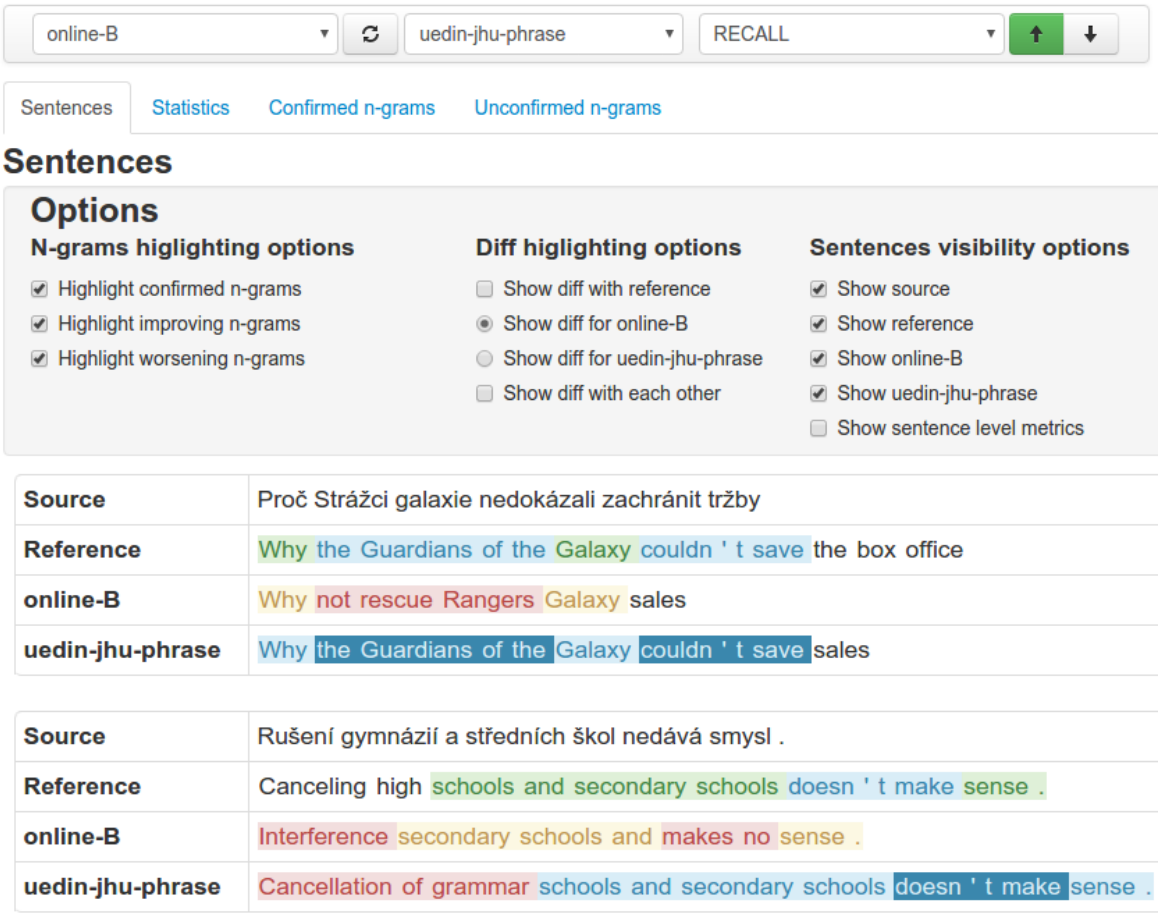


Figure 3: Online-B shortens translations.

- **Improving n-grams** are confirmed n-grams occurring in only one of the systems. These are highlighted in the system outputs with darker yellow and blue ("the Guardians of the" and "couldn't save" is present only in UEdin-jhu-phrase).

- **Worsening n-grams** are unconfirmed n-grams (i.e. probably wrong translations) occurring in only one of the systems. These are highlighted with red (e.g. "not rescue Rangers").

- **Diff** of the reference and one of the systems: words in the longest common subsequence of the two sentences can be underlined in green, other words in red – this was switched off in Figure 3 to keep it uncluttered.

Finding example sentences MT researchers often need to find a nice example where their system outperforms another system due to a given linguistic phenomenon. They can hide everything except for the colored reference translation (so more sentences fit one screen) and quickly search for a long enough blue-highlighted phrase exhibiting the phenomenon.

3.3. Statistics Pane

This pane focuses on quantitative evaluation and shows all document-level metric scores for the two systems compared and four area charts. The bottom two charts show *(non-paired) bootstrap resampling* (Koehn, 2004) for the two systems, to assess BLEU (or other selected metric) confidence intervals for the individual systems. We will focus on the two upper charts, depicted in Figure 4, where we compare Neural-MT and Online-A.

The left chart shows sentence-level BLEU-cis difference (y-axis) for all the 2656 sentences in the testset (x-axis): about half of the sentences are translated better by Neural-MT (green region) and half by Online-A (red region). Even if the red and green regions have the same area (which seems to be the case here), it does not imply that the document-level BLEU-cis are the same: document-level BLEU-cis is influenced more by longer sentences, moreover, it is not decomposable to sentence-level scores due to brevity penalty etc. (Chiang et al., 2008). Nevertheless, it is interesting to see what portion of sentences is better in one system with a given sentence-level BLEU margin compared with the other system.

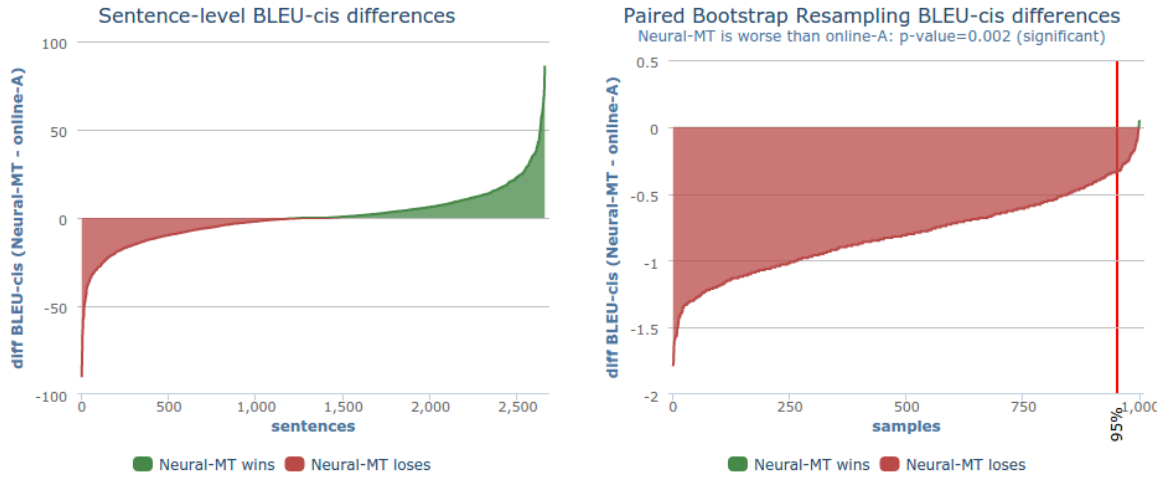


Figure 4: Statistics pane comparing Neural-MT and Online-A systems using Sentence-level BLEU differences graph (left) and Paired bootstrap resampling BLEU graph (right).

Significance MT researchers often need to know whether the difference between two systems in a given metric is significant or not. If the confidence intervals for the individual systems (in the bottom charts, not shown here) are not overlapping, it implies a significant difference, but the opposite implication does not hold. We need to use a paired test for checking the significance.

The right chart in Figure 4 shows *paired bootstrap resampling* (Koehn, 2004), where the x-axis lists 1,000 resamples of the testset and the y-axis is the difference in (document-level) BLEU-cis between the two systems for the given resample. One-tailed p-value is reported in the chart header: $p = 0.002$. This means that in 2 cases out of the 1000 resamples, Neural-MT had higher BLEU-cis than Online-A (this corresponds to the tiny green area on right, above the zero line). In the remaining 998 cases, Online-A had better scores, so we can conclude that Online-A is significantly better than Neural-MT in BLEU-cis (on the standard 95% confidence level).[8] When we change the metric to BLEU (case sensitive variant), we can see that Online-A is still better, but insignificantly ($p = 0.415$).

3.4. Confirmed and Unconfirmed n-grams Panes

Figure 5 shows Confirmed and Unconfirmed n-grams panes, focusing on unigrams only and merging the two panes into one figure for space reasons. We selected TectoMT and UEdin-jhu-phrase (the worst and the best system according to BLEU) for comparison. After clicking on any n-gram, a Sentence pane is opened showing all sentences with this n-gram (which is highlighted).

[8]For better reliability, the number of resamples can be increased in the configuration file, in option `bootstrapSampler`. MT-ComparEval uses 1000 resamples by default in order to import quickly new tasks. It also uses a random seed, so replicating the experiment may lead to slightly different p-values, e.g. the current version at `wmt.ufal.cz` has $p = 0.004$.

Quotes style In the second row of the table of confirmed unigrams where UEdin-jhu-phrase "wins", we can see (vertical) double quotes with numbers $590 - 188 = 402$. This means that this token was present 590 times in UEdin-jhu-phrase and confirmed by the reference, while TectoMT had only 188 confirmed occurrences of this token. The n-grams in the table are sorted according to the difference in the number of confirmed occurrences. In the 8th row of the table of unconfirmed unigrams where TectoMT loses, we can see (typographic) lower double quotes with numbers $247 - 0 = 247$. Lower quotes are used in Czech as opening quotes, but they should not be present in the English translations. The 0 means that UEdin-jhu-phrase did not produced any lower quotes (unconfirmed by the reference). TectoMT had 247 such occurrences and also 225 occurrences of unconfirmed (typographic) upper quotes. Thanks to MT-ComparEval, TectoMT developers were able to detect this error and fix it (simply by `s/[„“]/"/g`). In a similar way, un/confirmed n-grams are useful for quick spotting of various encoding problems (which may be more important for the translation quality than quotes style).

Definite and indefinite articles Figure 5 also reveals a problem with articles in TectoMT output. Table 1 summarizes the relevant numbers and computes the total number of occurrences of "the" and "a" in the systems' outputs.

We see that UEdin-jhu-phrase better uses "the" (confirmedDiff $= 2076 - 894 = 1182$), and it may seem that TectoMT better uses "a" (confirmedDiff $= 655 - 575 = 80$). It is important to always check also the Unconfirmed n-grams to prevent misleading conclusions. In Unconfirmed n-grams, we see that the seeming strengths of the systems are also their weaknesses: UEdin-jhu-phrase has 949 more unconfirmed "the"s than TectoMT, and TectoMT has 1065 more unconfirmed "a"s than UEdin-jhu-phrase.

We conclude that a) TectoMT produces in total fewer articles than UEdin-jhu-phrase. b) TectoMT prefers "a", while UEdin-jhu-phrase prefers "the". c) For "the", TectoMT has higher precision than UEdin-jhu-phrase; for "a" vice versa

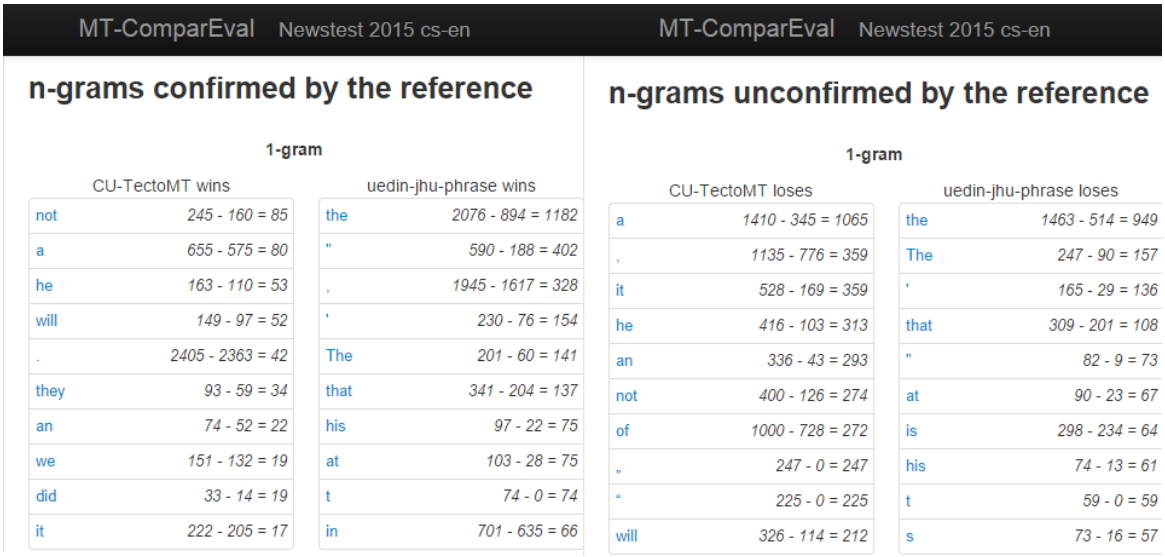


Figure 5: Confirmed and Unconfirmed n-grams panes (showing problems with quotes and articles in TectoMT).

	“the”		“a”	
	UEdin	TectoMT	UEdin	TectoMT
confirmed	2076	894	575	655
unconfirmed	1463	514	345	1410
total	3539	1408	920	2065
% confirmed	59%	63%	63%	32%

Table 1: Comparison of “the” and “a” usage in TectoMT and UEdin-jhu-phrase.

(see the last row in Table 1). d) Based on the number of confirmed n-grams, we see that for “the”, UEdin-jhu-phrase has much higher recall than TectoMT; for “a”, TectoMT has slightly higher recall than UEdin-jhu-phrase. e) With regards to the precision-recall balancing, TectoMT should produce more definite articles, but fewer indefinite ones.

Untranslated chunks Now, we will focus on Neural-MT and compare it with UEdin-jhu-phrase. Figure 6 shows unconfirmed unigrams and in the “Neural-MT loses” table, we can see “na” and “se”, which are often erroneously produced by Neural-MT (58 and 57 times, respectively), but never by UEdin-jhu-phrase. These tokens are frequent Czech words (prepositions).[9] If we click on these tokens, we will see sentences where Neural-MT left untranslated these tokens, quite often within longer untranslated phrases. For example, in Figure 7 we see untranslated phrases “První jarní den” (first day of spring) and “na letišti na letišti” (on airport on airport). Also in many other sentences found with “se”, “na” or “v”, we can see untranslated phrases consisting of easy-to-translate common words. We hypothesize that this is a peculiarity related to recurrent-neural-network

[9]Also “s” and “v”, which are listed in the table, are Czech prepositions, but these are sometimes erroneously produced also by UEdin-jhu-phrase. Due to the tokenization in MT-ComparEval (taken from BLEU), “he’s” is tokenized as “he ’ s” and thus token “s” may appear in English translations.

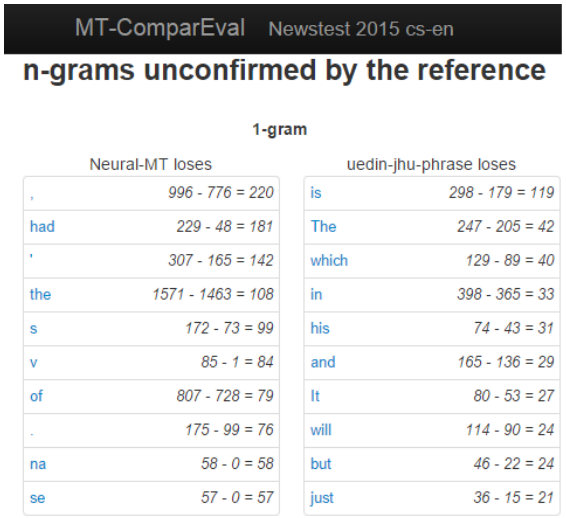


Figure 6: Untranslated Czech prepositions in Neural-MT.

nature of Neural-MT (Jean et al., 2015), which could be easily fixed (at least with an automatic post-processing).

Other Neural-MT peculiarities We noticed that the English translations contain not only untranslated Czech phrases, but also Czech phrases which were not in the source sentence, e.g. “které byly” (which were) in Figure 7. We also noticed many mistakenly repeated words or phrases (both translated and untranslated), e.g. “na letišti” (on airport). MT-ComparEval does not have any specialized tool for finding such repeated phrases, but the red highlighting in Sentence pane helps to spot them. Also the top unconfirmed Neural-MT 4-gram is “. . . .”, originating from a translation with a dot repeated 59 times.

Source	První jarní den poznamenán deštěm a bouřkami , které měly dopad na let na letišti v Adelaide
Reference	First day of spring marked with wet and blustery conditions impacting Adelaide Airport flights
Neural-MT	První jarní den by rain a storms , které byly impact na letišti na letišti v Adelaide

Figure 7: Example of Neural-MT output with untranslated phrases and unconfirmed unigram "na" highlighted.

4. Conclusion

In this paper, we have presented MT-ComparEval, an open-source tool that provides a graphically rich environment to perform quantitative and qualitative evaluation and deep analysis of machine translation outputs. We have presented its usage in the comparison and improvement of several systems.

While the developers of the underlying MT systems may already be familiar with many of the issues in their systems' output, MT-ComparEval helps to integrate quantitative analyses including significance tests with qualitative analysis that can help to avoid the most frequent systematic errors. This is especially relevant when working on "difficult" languages where fixing issues can be very costly, and thus has to be prioritized and systematic.

We are convinced that the usage of tools like MT-ComparEval in general will lead to a more analytic approach to MT development and evaluation, getting away from the very superficial level of "System A is better than system B by 1.5 BLEU score". It will help researchers to generate informed hypotheses for improvements and to increase the informativeness of publications as the graphical interface makes it easy to search for nice illustrating examples that fix certain issues under consideration (or lead to new issues to be fixed).

5. Acknowledgment

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), GA15-10472S (Manyla), SVV 260 224, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071). We thank the three anonymous reviewers for useful comments.

6. Bibliographical References

Agarwal, A. and Lavie, A. (2008). Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal, September. Association for Computational Linguistics.

Chiang, D., DeNeefe, S., Chan, Y. S., and Ng, H. T. (2008). Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619, Honolulu, Hawaii, October. Association for Computational Linguistics.

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. (2015). Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September. Association for Computational Linguistics.

Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-ComparEval: graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, (104):63–74.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Dekang Lin et al., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. ACL.

Koehn, P. (2010). An Experimental Management System. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Stroudsburg, PA, USA. ACL.

Popović, M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, pages 59–68.