MT-TELESCOPE
An interactive platform for contrastive evaluation of MT systems

Ricardo Rei[*†‡] **Ana C Farinha**[*] **Craig Stewart**[*]
Luisa Coheur[†‡] **Alon Lavie**[*]
[*]Unbabel Research
[‡]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[†]INESC-ID, Lisboa, Portugal
[*]{firstname.lastname}@unbabel.com
luisa.coheur@inesc-id.pt

Abstract

We present MT-TELESCOPE, a visualization platform designed to facilitate comparative analysis of the output quality of two Machine Translation (MT) systems. While automated MT evaluation metrics are commonly used to evaluate MT systems at a corpus-level, our platform supports fine-grained segment-level analysis and interactive visualisations that expose the fundamental differences in the performance of the compared systems. MT-TELESCOPE also supports dynamic corpus filtering to enable focused analysis on specific phenomena such as; translation of named entities, handling of terminology, and the impact of input segment length on translation quality. Furthermore, the platform provides a bootstrapped t-test for statistical significance as a means of evaluating the rigor of the resulting system ranking. MT-TELESCOPE is open source[1], written in Python, and is built around a user friendly and dynamic web interface. Complementing other existing tools, our platform is designed to facilitate and promote the broader adoption of more rigorous analysis practices in the evaluation of MT quality.

1 Introduction

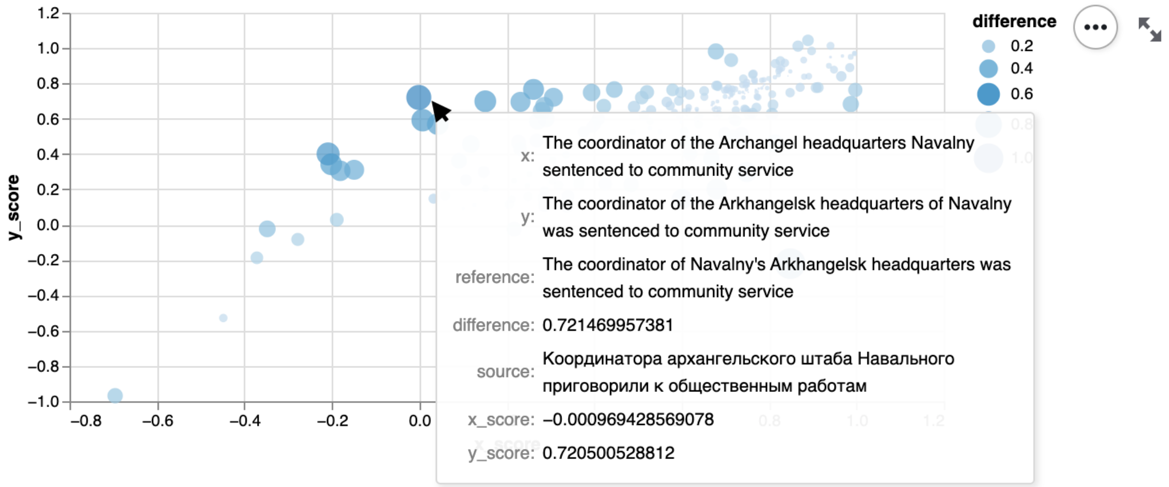
When developing MT systems or comparing experiments across papers, it has been common practice for researchers and developers to rely on automated metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) as a means of quantifying the relative performance difference between two models. Commercial deployment of systems and the establishment of state-of-the-art in academia is often driven by these metrics alone. Automated metrics have long been an essential means for assessing quality improvements and driving progress in the field of MT. Recent state-of-the-art metrics such as COMET (Rei et al., 2020a), PRISM (Thompson and Post, 2020), and BLEURT (Sellam et al., 2020), show much higher levels of correlation with human judgement than their predecessors.

Notwithstanding the strength of available metrics, when applied and reported at corpus-level, they are only able to provide a general indication of whether one system is superior, based on a single score which in some cases is limited to an arithmetic mean of segment-level score predictions (Rei et al., 2020a). We contend that the broad definition of 'improvement' as an increase in a relevant corpus-level score is insufficient, especially when the relative difference between high-performing MT systems is negligible. Exposure of the changing distribution of performance at segment-level on targeted phenomena is fundamental to our understanding of translation quality. Manual inspection at this level is often too time-consuming and inefficient to be done rigorously and on a regular basis.

MT-TELESCOPE was inspired by other recent work on developing holistic approaches for fine-grained comparison of MT systems, such as COMPARE-MT (Neubig et al., 2019) and MT-COMPAREVAL (Klejch et al., 2015) and other more general comparative tools such as VIZSEQ (Wang et al., 2019). Despite the intention of such tools in addressing the above problem, none have been widely adopted as a standard method of evaluating MT. MT-TELESCOPE was specifically developed to leverage the best of existing approaches in a manner that is as user friendly as possible, with features specifically tailored to the MT use case. The platform supports fine-grained segment-level analysis and interactive visualisations that provide relevant and informative quality intelligence. In particular, the platform also supports focused anal-

[1] Code available at: https://github.com/Unbabel/MT-Telescope and Demo video at: https://youtu.be/MZOe1yX8mII

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 73–80, August 1st - August 6th, 2021.



x: The coordinator of the Archangel headquarters Navalny sentenced to community service

y: The coordinator of the Arkhangelsk headquarters of Navalny was sentenced to community service

reference: The coordinator of Navalny's Arkhangelsk headquarters was sentenced to community service

difference: 0.721469957381

source: Координатора архангельского штаба Навального приговорили к общественным работам

x_score: −0.000969428569078

y_score: 0.720500528812

Figure 1: Segment comparison bubble plot.

ysis of MT-specific phenomena through interactive corpus filtering.

MT-TELESCOPE is differentiated from existing MT-specific tools by exposing features such as named entities and glossary handling which play a fundamental role in determining the suitability of an MT system for a production environment. Furthermore, the platform applies a bootstrapped t-test for statistical significance (Koehn, 2004) as a means of exposing the experimental rigor of system comparisons. These features are not widely available in other tools and provide a uniquely tailored solution to MT comparison that is highly informative and easy to use.

The fundamental goal of MT-TELESCOPE is to widen access to state-of-art, robust MT comparison, to the benefit of the MT community at large. MT-TELESCOPE is open source, written in Python and uses a dynamic web interface implemented in streamlit[2]. In this manner, MT-TELESCOPE provides a uniquely accessible framework that requires little technical skill to operate and exposes information about the critical differences between MT outputs that is interactive, informative and highly customizable.

2 MT-TELESCOPE: Features

In this section, we describe the main features and visualizations implemented in MT-TELESCOPE and illustrate the user experience with examples:

2.1 User input and data

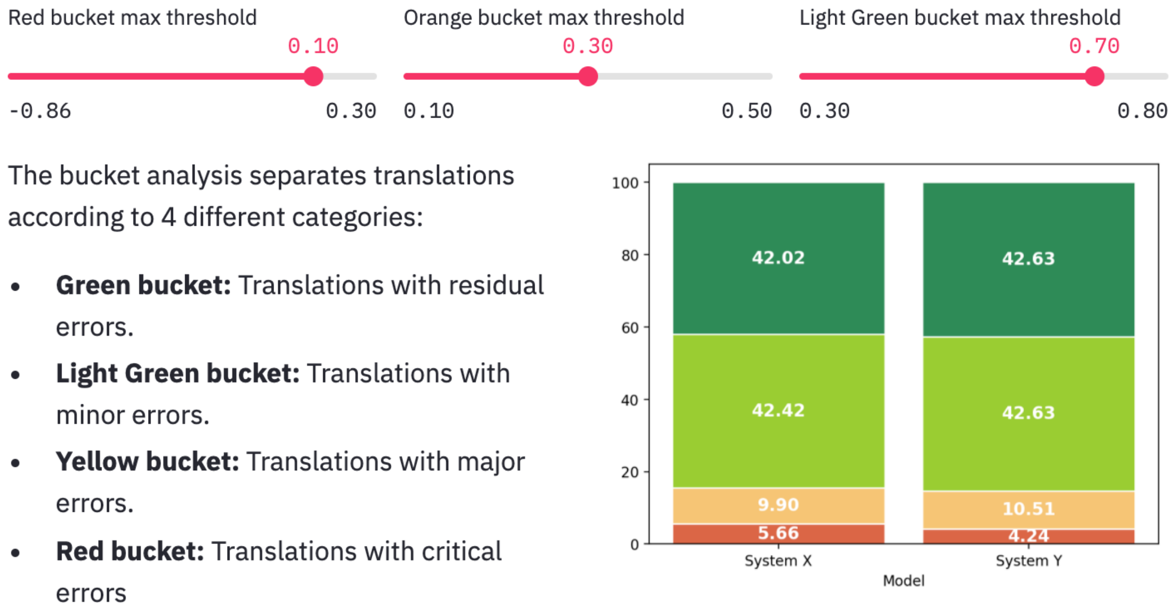
MT-TELESCOPE is opened in a web browser and takes four text (*.txt*) files as input; source and reference segments and one set of MT outputs for each of the compared systems. Users drag and drop these files directly onto the interface to begin evaluation. COMET (Rei et al., 2020a) is provided as a default metric given its proven value in the WMT Metrics Shared Task 2020 (Rei et al., 2020b; Mathur et al., 2020). Optionally the user can choose an alternate metric using a selection box. Currently available metrics include BLEU, METEOR and CHRF, and a selection of more recently proposed metrics such as PRISM, BLEURT, and BERTSCORE.

2.2 Visualizations

High-level results of the analysis are output in table format with the corresponding system scores. MT-TELESCOPE then exposes segment-level comparison in three primary visualizations:

First, a bubble plot (Figure 1) where the position of bubbles show how scores between the two systems differ for each segment, notable differences being highlighted with variations in bubble size and color. This method of visualization of MT is unique to MT-TELESCOPE in that it is fully interactive; by hovering the cursor over individual data points the user can preview the segments and output as well as relevant scores and the magnitude of the difference between them (as depicted in Figure 1). This plot allows for interactive exploration of the data which easily exposes differences in model

[2] https://streamlit.io/



Red bucket max threshold 0.10 (-0.86 – 0.30)

Orange bucket max threshold 0.30 (0.10 – 0.50)

Light Green bucket max threshold 0.70 (0.30 – 0.80)

The bucket analysis separates translations according to 4 different categories:

- **Green bucket:** Translations with residual errors.
- **Light Green bucket:** Translations with minor errors.
- **Yellow bucket:** Translations with major errors.
- **Red bucket:** Translations with critical errors

Figure 2: Segment-level error bucket analysis plot. In this plot, we can compare the two systems side by side according to the percentage of segments falling into 4 different category buckets: *residual errors*, *minor errors*, *major errors*, *critical errors*. The thresholds for defining these buckets can be dynamically adjusted using the sliders displayed above the plot.

behaviour at a glance. In particular, the distribution of points along the diagonal of this plot is highly informative; clustering along the diagonal indicates that the systems have minor differences whereas the contrary can indicate more dramatic change in behavior which can be hidden by the corpus-level mean.

Second, MT-TELESCOPE provides a bucketed error analysis in the form of a stacked bar plot (Figure 2). This plot serves to isolate specific bands of translation quality. These bands are highly customizable but can serve as a means of evaluating system utility; the plot can expose the extent to which either model outputs critical error for example. This is particularly useful in a commercial setting where the utility of a production system is inhibited by the presence of particular error types.

Segments are grouped into four buckets: *residual errors*, *minor errors*, *major errors*, and *critical errors*. The thresholds for each bucket can be dynamically adjusted by the user with appropriate sliders and (as with many of the features of MT-TELESCOPE) the plots are updated in real-time to reflect adjustments. Defaults were determined in line with suggestions outlined in the COMET GitHub documentation and with distributions of system-level scores from the WMT News Translation Shared Task 2020.

Residual Errors: The highest tier of quality by default reflects scores greater than 0.70, which generally equates to almost human-like translation with only minor, inconsequential error.

Minor Errors: By default this band reflects scores between 0.30 and 0.70 to reflect the division of quartiles from the distribution of system-level scores from the WMT News Translation Shared Task 2020. In general the band is associated with translation that is adequate but with minor flaws.

Major Errors: Translations scoring between 0.10 and 0.30 by default inhabit this band and are generally inadequate due to more serious error.

Critical Errors: Any translation scoring under 0.10 here is considered to contain critical error.

These bands are intended as a guide and utility of the default thresholds will vary according to use case. Translation quality and the difference between adequate and inadequate translation is highly subjective and language dependant; optimization of these thresholds is a critical direction for future work. Notwithstanding, we find that exposure of the general shift in distribution of inadequate translation in general is potentially informative, particularly given that corpus-level scores do not expose this type of analysis.

Finally, MT-TELESCOPE provides a histogram plot (Figure 3) for general evaluation of the distri-

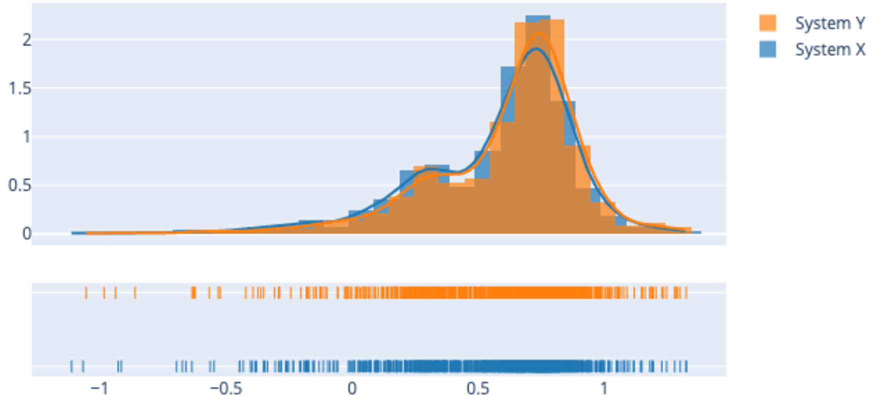


Figure 3: Segment-level histogram comparison.

bution of scores between models. We propose that this kind of plot can potentially provide a high-level overview of the shift in performance between models. A corpus-level score (particularly an arithmetic mean) can mask variance between distributions of scores.

2.3 Example evaluation

To demonstrate the utility of the MT-TELESCOPE evaluation we expose analyses for the *Online-G* and the *PROMT* (Molchanov, 2020) systems from the WMT News Translation Shared Task 2020 (Barrault et al., 2020) for Russian-English:

The *Online-G* system (System Y) achieves a COMET score of 0.6081, outperforming the *PROMT* system (System X) which only achieves 0.5972. We have isolated this example in particular as it represents a common occurrence of two systems achieving fairly comparable scores.

Figures 1, 2 and 3 above show the output of MT-TELESCOPE analysis on two sampled systems:

Figures 2 and 3 illustrate that the second system (System Y) in general exceeds performance of the first (System X). We can conclude from these plots that the systems perform comparably with System Y producing a higher percentage of adequate translations. In particular we note that System Y outputs fewer critical errors, consistent with its general performance gain.

Figure 1 illustrates isolation of an example where System Y makes substantial gain over System X. Here we note that both systems struggle to render the named entity and the corresponding possessive, but that System Y successfully produces the named entity as reflected in the reference and adds a pronoun to at least give possessive flavor.

3 MT-TELESCOPE: Dynamic Corpus Filtering

Given a test corpus, MT-TELESCOPE provides functionality to dynamically evaluate sub-samples of the system outputs as a means of focused analysis tailored to particular phenomena relevant to MT. On selection of any of the available filtering criteria, the MT-TELESCOPE Dynamic Corpus Filtering feature (DCF) updates the output evaluation in real-time to allow the user to ‘zoom in’ on relevant data points.

Currently, MT-TELESCOPE supports filtering by named entity, glossary and source segment length, as well as an option to remove duplicates. Whenever any of these options is selected, the interface will output the size of the sub-sample as a percentage of the original test corpus.

3.1 DCF: Named Entities

Successful rendering of named entities is a known challenge for even modern MT systems and can lead to distortion of locations, organization and other names (Koehn and Knowles, 2017; Modrzejewski et al., 2020). Recently, several methods have been proposed to improve the translation

Table 1: Example of named entity errors produced *Online-G* system in comparison to the *PROMT* system from the WMT20 shared task.

		Comet
Source	Маругов врезался на мотоцикле в такси, которым управлял Акбаров.	
Online-G	Murugov crashed into a motorcycle taxi, which was ruled by Akbar.	-0.1799
PROMT	Marugov crashed into a taxi driven by Akbarov on a motorcycle.	0.5154
Reference	Marugov crashed on a motorcyle into the taxi Akbarov was driving.	

of named entities in Neural Machine Translation (NMT) (Sennrich and Haddow, 2016; Ugawa et al., 2018; Modrzejewski et al., 2020), but precise measurement of translation quality improvements for these techniques is inhibited by the fact that not all sentences in traditional benchmark test sets (e.g. WMT test sets) contain named entities and that scores produced by automated evaluation metrics are not sufficiently fine-grained to reflect this type of variation. MT-Telescope offers a potential solution to this by applying the following filter:

We initially run the Stanza Named Entity Recognition (NER) model (Stanza, Qi et al. 2020)[3] over the source test corpus to isolate segments that contain named entities. If the source language (as specified by the user) is not supported by Stanza, we run NER on the reference. MT-Telescope will then update the output analysis allowing focused evaluation of the handling of segments containing named entities by either MT system.

To illustrate the utility of DCF analysis on named entities we again compare the outputs of the *Online-G* and the *PROMT* (Molchanov, 2020) systems from the Metrics Shared Task 2020 (Barrault et al., 2020) as above:

Applying DCF for named entities, the *Online-G* system Comet score drops to 0.5851 (previously 0.6081), while the *PROMT* system only drops to 0.5888 (previously 0.5972). We also observe that the percentage of critical segments from the *Online-G* system in our bucketed analysis jumps from 6.26% to 7.0%, while the corresponding percentage output by the *PROMT* system drops from 6.66% to 6.29%.

On the basis of the DCF analysis for named entities we can conclude that whilst in general the *Online-G* exhibits superior quality, it may be underperforming with regard to named entities. Interestingly, the system description paper for the *PROMT* system (Molchanov, 2020) specifically details a targeted approach to handling translation of named entities, which may explain its stronger performance on the isolated sub-sample.

In Table 1 we illustrate an example of a translation in which the *Online-G* system produces critical errors as a consequence of translating named entities incorrectly, specifically isolated by the DCF feature.

3.2 DCF: Terminology

Similarly to named entities, enforcing that MT systems use specific terminology during translation is a challenging task with particular relevance in commercial use cases. Measuring terminology adherence typically involves relying on automated metrics for MT as well as measuring the accuracy of terminology output (Dinu et al., 2019; Exel et al., 2020).

This approach presents two concrete problems: a) applying terminology constraints typically results in only minimal variance between translations, which limits the utility of using automated metrics at the corpus level; and b) measuring accuracy in terminology usage typically relies on exact string matching between a translation hypothesis and its respective reference, which implies that properly inflected translated terms often do not receive proper credit.

MT-Telescope offers a DCF Terminology feature which allows a user to optionally upload a glossary by which to isolate a corresponding subsample of the test corpus. We apply string matching on the source and filter to only those segments which contain a corresponding glossary match.

3.3 DCF: Segment Length

Another common weakness of some MT systems is their inability to accurately translate long segments (Koehn and Knowles, 2017). In general, corpus level evaluation on a distribution that includes very short segments can artificially inflate performance, with substantial drops in scores being observed when these segments are specifically excluded (Koehn and Knowles, 2017). In the same manner, quality-based decisions regarding two systems can change when we consider segments of

[3] https://stanfordnlp.github.io/stanza/ner.html

different lengths.

Using our example systems outlined above in Section 3.1, when comparing the *Online-G* and the *PROMT* systems using only the top 50% longest segments, the *PROMT* system outperforms the *Online-G* system according to COMET and CHRF scores, changing the fundamental perception of which system is 'better'. With the above in mind, MT-TELESCOPE also offers an option to filter by segment length. This filter is adaptive to the distribution of segment lengths in the test corpus. We first build the distribution of the source segment lengths (measured in terms of characters) for the entire test set. Then, the user can select which part of the distribution to analyse by adjusting the a and b parameters of the density function $P(a \leq X \leq b)$; a and b being the minimum and maximum length allowed, respectively.

3.4 DCF: Duplication

The removal of duplicates can be particularly important in situations where the test corpus sample contains repetition. Repeated segments in a test sample can artificially inflate the corpus-level score, particularly where that score results from an average of segment-level scores. Whilst we acknowledge that removal of duplicate segments is fairly common in public data sets such as that used in the WMT Shared Tasks and consequently our example here, we propose that it is, nevertheless, a useful tool when evaluating on random samples.

4 Statistical Significance Testing

By default, MT-TELESCOPE implements the bootstrapped t-test for statistical significance promoted for use in comparison of MT systems by Koehn (2004). Specifically, we iteratively re-sample a portion of the test set (of size P) N times, compare corpus-level results of each sub-sample and record the comparative conclusions. The ratio of wins of a single system is a reasonable proxy to the probability that that system is better than the other. In other words, if one system outperforms the other system 95% of the time, we conclude that the former is better with a significance of $p = 0.05$ (Koehn, 2004).

This is particularly useful in cases where the relative difference between systems is minimal and acts as a measure of the robustness of any resulting decision. In our implementation P is an optional parameter which defaults to 0.5 (50%) or 500 segments, whichever is larger, to ensure reasonable stability in the output conclusion. N is also user defined and by default is set at 300 iterations.

5 Related Tools

MT-TELESCOPE is similar in spirit and largely inspired by recently proposed tools such as COMPARE-MT (Neubig et al., 2019), MT-COMPAREVAL (Klejch et al., 2015), and VIZSEQ (Wang et al., 2019). COMPARE-MT also provides a holistic analysis comparing two MT systems, although with different features. Using COMPARE-MT, the user can, for example, look at performance according to n-gram frequency and part-of-speech (POS) accuracy. MT-COMPAREVAL also provides comparative analysis of segment-level errors with highlighting of variant n-grams. The tool also provides some limited aggregate analysis. Both of the above tools also offer statistical significance testing in the form of a bootstrapped t-test.

VIZSEQ (Wang et al., 2019), whilst only tangentially related, is one of the only comparative tools that offers a web-based interface. Moreover, VIZSEQ has impressive coverage in terms of Natural Language Generation metrics. However, VIZSEQ was developed for multi-model comparison and is primarily focused at corpus-level. Other tools such as PET (Aziz et al., 2012) and APPRAISE (Federmann, 2012) are complementary to MT-TELESCOPE in that they offer features which leverage annotation and post-edition.

6 Conclusions and Future Work

MT-TELESCOPE is designed to provide robust and insightful comparative analysis specific to the MT use case with state-of-the-art metrics. Data visualizations are dynamic, interactive and highly customizable. The tools have been built specifically with ease of use in mind, in the hope of expanding access to high quality MT evaluation.

There is tremendous scope in the adaptation of the DCF framework to target many other phenomena and future work will be focused primarily in this area. We envisage for example adding filters for specific discourse phenomenon such as pronoun translation. Ideally such filter would allow researchers to measure context usage in NMT without having to rely only on contrastive evaluation (Müller et al., 2018; Lopes et al., 2020) and/or human evaluation.

We also plan to extend MT-TELESCOPE to handle a (possibly empty) set of references. This will bring more flexibility to the tool allowing more informed decision when multiple references are available while also supporting Quality Estimation (Specia et al., 2018) when references are not available. Finally we hope to implement exporting functionality to allow saving of analysis output in commonly used formats (e.g. json and PDF). Given that MT-TELESCOPE is an open source platform, we are excited to encourage other users to contribute to its growth with suggestions and new features.

Acknowledgments

We are grateful to the Unbabel MT team, specially Austin Matthews and João Alves, for their valuable feedback. This work was supported in part by the P2020 Program through projects MAIA and Unbabel4EU, supervised by ANI under contract numbers 045909 and 042671, respectively.

References

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA).

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for Machine Translation development. *The Prague Bulletin of Mathematical Linguistics*, 104.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Alexander Molchanov. 2020. PROMT systems for WMT 2020 shared news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 248–253, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. VizSeq: a visual analysis toolkit for text generation tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 253–258, Hong Kong, China. Association for Computational Linguistics.