Vis-Eval Metric Viewer: A Visualisation Tool for Inspecting and Evaluating Metric Scores of Machine Translation Output

David Steele and **Lucia Specia**
Department of Computer Science
University of Sheffield
Sheffield, UK
dbsteele1,l.specia@sheffield.ac.uk

Abstract

Machine Translation systems are usually evaluated and compared using automated evaluation metrics such as BLEU and METEOR to score the generated translations against human translations. However, the interaction with the output from the metrics is relatively limited and results are commonly a single score along with a few additional statistics. Whilst this may be enough for system comparison it does not provide much useful feedback or a means for inspecting translations and their respective scores. Vis-Eval Metric Viewer (VEMV) is a tool designed to provide visualisation of multiple evaluation scores so they can be easily interpreted by a user. VEMV takes in the source, reference, and hypothesis files as parameters, and scores the hypotheses using several popular evaluation metrics simultaneously. Scores are produced at both the sentence and dataset level and results are written locally to a series of HTML files that can be viewed on a web browser. The individual scored sentences can easily be inspected using powerful search and selection functions and results can be visualised with graphical representations of the scores and distributions.

1 Introduction

Automatic evaluation of Machine Translation (MT) hypotheses is key for system development and comparison. Even though human assessment ultimately provides more reliable and insightful information, automatic evaluation is faster, cheaper, and often considered more consistent.

Many metrics have been proposed for MT that compare system translations against human references, with the most popular being BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and, more recently, BEER (Stanojevic and Sima'an, 2014). These and other automatic metrics are often criticised for providing scores that can be non-intuitive and uninformative, especially at the sentence level (Zhang et al., 2004; Song et al., 2013; Babych, 2014). Additionally, scores across different metrics can be inconsistent with each other. This inconsistency can be an indicator of linguistic properties of the translations which should be further analysed. However, multiple metrics are not always used and any discrepancies among them tend to be ignored.

Vis-Eval Metric Viewer (VEMV) was developed as a tool bearing in mind the aforementioned issues. It enables rapid evaluation of MT output, currently employing up to eight popular metrics. Results can be easily inspected (using a typical web browser) especially at the segment level, with each sentence (source, reference, and hypothesis) clearly presented in interactive score tables, along with informative statistical graphs. No server or internet connection is required. Only readily available packages or libraries are used locally.

Ultimately VEMV is an accessible utility that can be run quickly and easily on all the main platforms.

Before describing the technical specification of the VEMV tool and its features in Section 3, we give an overview of existing metric visualisation tools in Section 2.

2 Related Tools

Several tools have been developed to visualise the output of MT evaluation metrics that go beyond displaying just single scores and/or a few statistics.

Despite its criticisms and limitations, BLEU is still regarded as the *de facto* evaluation metric used for rating and comparing MT systems. It was one of the earliest metrics to assert a high enough correlation with human judgments.

Interactive BLEU (iBleu) (Madnani, 2011) is

Proceedings of NAACL-HLT 2018: Demonstrations, pages 71–75
New Orleans, Louisiana, June 2 - 4, 2018. ©2018 Association for Computational Linguistics

a visual and interactive scoring environment that uses BLEU. Users select the source, reference, and hypothesis files using a graphical user interface (GUI) and these are scored. The dataset BLEU score is shown alongside a bar chart of sentence scores. Users can select one of the sentences by clicking on the individual bars in the chart. When a sentence is selected its source and hypothesis translation is also shown, along with the standard BLEU statistics (e.g. score and n-gram information for the segment). Whilst iBLEU does provide some interactivity, using the graph itself to choose the sentences is not very intuitive. In addition the tool provides results for only one metric.

METEOR is another popular metric used to compute sentence and dataset-level scores based on reference and hypothesis files. One of its main components is to word-align the words in the reference and hypothesis. The Meteor-X-Ray tool generates graphical output with visualisation of word alignments and scores. The alignments and score distributions are used to generate simple graphs (output to PDF). Whilst the graphs do provide extra information there is little in the way of interactivity.

MT-ComparEval (Klejch et al., 2015) is a different evaluation visualisation tool, available to be used online[1] or downloaded locally. Its primary function is to enable users, via a GUI, to compare two (or more) MT system outputs, using BLEU as the evaluation metric. It shows results at both the sentence and dataset level highlighting confirmed, improving, and worsening n-grams for each MT system with respect to the other. Sentence-level metrics (also n-gram) include precision, recall, and F-Measure information as well as score differences between MT systems for a given sentence. Users can upload their own datasets to view sentence-level and dataset scores, albeit with a very limited choice of metrics. The GUI provides some interaction with the evaluation results and users can make a number of preference selections via check boxes.

The Asiya Toolkit (Giménez et al., 2010) is a visualisation tool that can be used online or as a stand-alone tool. It offers a comprehensive suite of metrics, including many linguistically motivated ones. Unless the goal is to run a large number of metrics, the download version is not very practical. It relies on many external tools such as syntactic and semantic parsers. The online tool[2] aims to offer a more practical solution, where users can upload their translations. The tool offers a module for sentence-level inspection through interactive tables. Some basic dataset-level graphs are also displayed and can be used to compare system scores.

In comparison to the other software described here, VEMV is a light yet powerful utility, which offers a wide enough range of metrics and can be easily extended to add other metrics. It has a very specific purpose in that it is designed for rapid and simple use locally, without the need for servers, access to the internet, uploads, or large installs. Users can quickly get evaluation scores from a number of mainstream metrics and view them immediately in easily navigable interactive score tables. We contend that currently there is no other similar tool that is lightweight and offers this functionality and simplicity.

3 Vis-Eval Metric Viewer Software & Features

This section provides an overview of the VEMV software and outlines the required input parameters, technical specifications, and highlights a number of the useful features.

3.1 The Software

VEMV is essentially a multi-metric evaluation tool that uses three tokenised text files (source, reference, and hypothesis) as input parameters and scores the hypothesis translation (MT system output) using up to eight popular metrics: BLEU, MT-Eval[3] (MT NIST & MT BLEU), METEOR, BEER, TER, Word Error Rate (WER), and Edit Distance (E-Dist).[4] All results are displayed via easily navigable web pages that include details of all sentences and scores (shown in interactive score tables - Figure 1). A number of graphs showing various score distributions are also created.

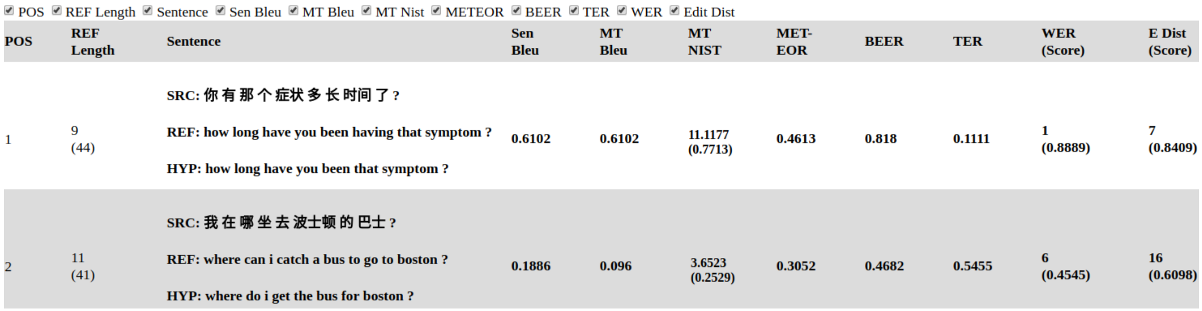
The key aims of VEMV are to make the evaluation of MT system translations easy to undertake and to provide a wide range of feedback that helps the user to inspect how well their system performed, both at the sentence and dataset level.

[1] http://wmt.ufal.cz/

[2] At the time of writing the online version did not work.

[3] https://www.nist.gov/

[4] A WER like metric that calculates the Levenshtein (edit) distance between two strings, but at the character level.



☑ POS ☑ REF Length ☑ Sentence ☑ Sen Bleu ☑ MT Bleu ☑ MT Nist ☑ METEOR ☑ BEER ☑ TER ☑ WER ☑ Edit Dist

POS	REF Length	Sentence	Sen Bleu	MT Bleu	MT NIST	MET-EOR	BEER	TER	WER (Score)	E Dist (Score)
1	9 (44)	SRC: 你有那个症状多长时间了？ REF: how long have you been having that symptom ? HYP: how long have you been that symptom ?	0.6102	0.6102	11.1177 (0.7713)	0.4613	0.818	0.1111	1 (0.8889)	7 (0.8409)
2	11 (41)	SRC: 我在哪坐去波士顿的巴士？ REF: where can i catch a bus to go to boston ? HYP: where do i get the bus for boston ?	0.1886	0.096	3.6523 (0.2529)	0.3052	0.4682	0.5455	6 (0.4545)	16 (0.6098)

Figure 1: A screenshot of an interactive score table showing two example sentences and their respective scores.

3.2 Input and Technical Specification

VEMV is written in Python 3 (also compatible with Python 2.7). To run the tool, the following software needs to be installed:

- Python >= 2.7 (required)
- NLTK[5] >= 3.2.4 (required)
- Numpy (required)
- Matplotlib / Seaborn (optional - for graphs)
- Perl (optional - for MT BLEU, MT NIST)
- Java (optional - for METEOR, BEER, TER)

With the minimum required items installed the software will generate scores for standard BLEU, WER, and E-Dist. The optional items enable a user to run a wider range of metrics and produce nearly 200 graphs during evaluation.

The input commands to run the software can be typed directly into the command-line on any platform, or passed as arguments in an interactive development environment (IDE) such as Spyder.[6]

Once the software has been run (see Section 3.5), a folder containing all of the generated HTML, text, and image files is produced. A user will typically explore the output by opening the 'main.html' file in a browser (Chrome, Firefox, and Opera have been tested) and navigating it like with any (offline) website. The text files contain the output for the various metric scores and can be inspected in detail. The graphs are output as image files (PNGs), which are primarily viewed in the HTML pages, but can also be used separately for reports (e.g. Figure 3 in Section 3.4)

3.3 Main Features

Here we outline some key features of the Vis-Eval Metric Viewer tool:

Scoring with multiple evaluation metrics

Currently VEMV uses eight evaluation metrics to score individual sentences and the whole document. All results are shown side by side for comparison purposes and can be inspected at a granular level (Figure 1).

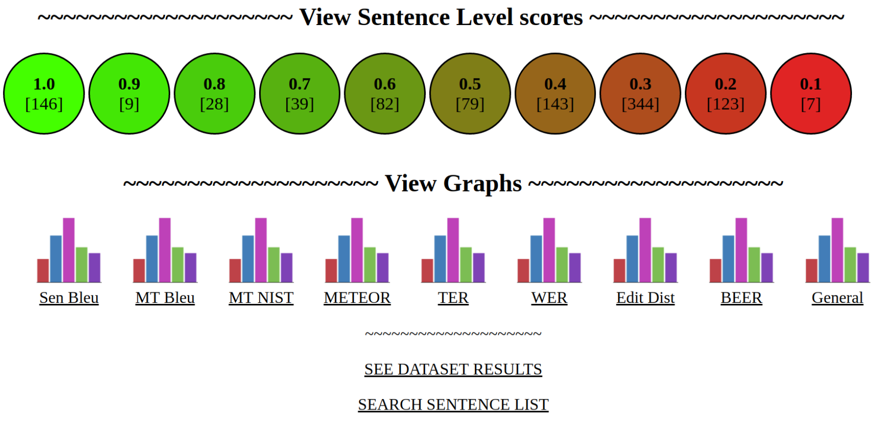
A glance at the two sentences in Figure 1 already provides numerous points for analysis. For example, the MT in sentence 2 is a long way from the reference and receives low metric scores. However, whilst not identical to the reference, the MT is correct and could be interchanged with the reference without losing meaning. For sentence 1 the MT is only a single word away from the reference and receives good scores, (much higher than sentence 2) although the meaning is incorrect. The interactive display enables the user to easily examine such phenomena in a given dataset.

Clear and easily navigable output

The main output is shown as a series of web pages and can be viewed in modern browsers. The browsers themselves also have a number of powerful built-in functions, such as page search, which are applicable to any of the output pages, adding an extra layer of functionality.

The output consists of easily navigable interactive score tables and graphs, logically organised across web pages. The tool includes its own search facility (for target and source sentences) and the option to show or hide metric scores to aid clarity, especially useful for comparing only a selection of metrics. All of the segment level metric scores can be sorted according to the metric of interest.

[5] http://www.nltk.org
[6] https://github.com/spyder-ide/spyder/releases



~~~~~~~~~~~~~~~~~~~~ View Sentence Level scores ~~~~~~~~~~~~~~~~~~~~

1.0 [146] 0.9 [9] 0.8 [28] 0.7 [39] 0.6 [82] 0.5 [79] 0.4 [143] 0.3 [344] 0.2 [123] 0.1 [7]

~~~~~~~~~~~~~~~~~~~~ View Graphs ~~~~~~~~~~~~~~~~~~~~

Sen Bleu MT Bleu MT NIST METEOR TER WER Edit Dist BEER General

~~~~~~~~~~~~~~~~~~~~

SEE DATASET RESULTS

SEARCH SENTENCE LIST

Figure 2: A screenshot of the VisEval Metric Viewer main page.

Results saved locally

Once scored, the generated text files, images, and HTML pages are saved locally in a number of organised folders. The majority of the text files are made up from the standard raw output of the metrics themselves. The image files are statistical graphs produced from the metric scores. Both the text and image files can be inspected directly on a metric by metric basis and used for reference. The VEMV tool brings together the text and images in the HTML files to form the main viewable output.

Runtime user options

The minimal default settings will quickly produce scores for standard BLEU, WER and E-Dist. Numerous parameters can be set on the command line enabling the user to choose any or all of the additional metrics and whether or not to generate graphs.

A number of the metrics (especially BLEU and METEOR) have a plethora of parameters, which can be selected. To avoid the need for complex command line inputs the metric level parameters can be placed in an easily editable text based configuration file, which in turn is passed to the command line.

In addition, the user can choose which metric will be the dominant one for sorting and display purposes (the default is BLEU) and there is an option for selecting how many score bins or pages to use to show the sentences. The default is 100 pages (one for every percentage point), but some users may prefer fewer pages (e.g. 10 or 20) in order to simplify the main interface and general navigation.

An accessibility flag has also been added. It removes some of the colour formatting from the displays making it easier for users with visual impairments (e.g colour blindness).

3.4 Viewing the Actual Output

Figure 2 shows the main page of the software. In this case all eight metrics were used as shown by the mini graph icons. Each of these mini graph icons act as a link. Ten score bins (circular icons) were selected as a parameter.

Users can click on any of the links/icons to navigate to the various pages. Clicking on the circular icons opens the sentence level score pages (Figure 1) showing individual sentences with a given score. Clicking on the mini graph icons takes the user to the graph display web pages for the respective metrics or the general document wide statistics. Figure 3, for example, is a metric graph showing the distribution of standard BLEU scores for the dataset. In this case the chart in Figure 3 would be accessed by clicking on the very left hand mini graph icon on the main page shown in Figure 2.

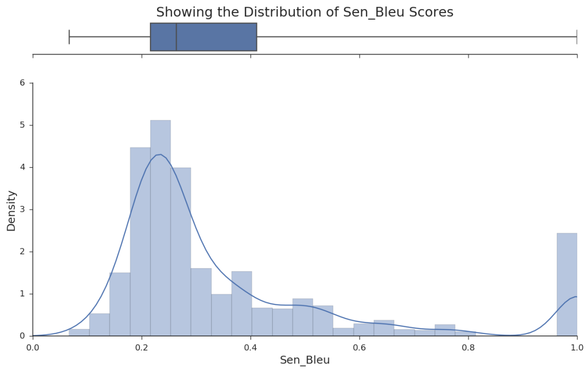


Figure 3: A graph showing the distribution of standard BLEU scores.
~~~~~~~~~~~~~~~~~~~~

3.5 Downloading and Running the Tool

Vis-Eval Metric Viewer can currently be downloaded from the following location on GitHub: https://github.com/David-Steele/VisEval_Metric_Viewer.

The associated README file provides instructions on how to get started with using the tool, and what to do if you run into any problems.

In terms of hardware requirements, a computer with at least 2GB of RAM and 300MB of available storage is needed to run the software.

A short video demonstration of these and other features of the Vis-Eval Metric Viewer software can be found online at: https://youtu.be/nUmdlXGYeMs.

4 Conclusion and Future Work

The Vis-Eval Metric Viewer tool was designed with three main aims:

- To provide a useful tool that is easy to install (using readily available packages), and simple to use and run on a local machine without the need for a server or internet connection.
- To offer a single place for scoring translations using multiple popular metrics.
- To provide in depth visual feedback making it easy to examine segment level metric scores.

The tool offers a light weight solution that makes it easy to compare multiple-metric scores in a clear manner. Feedback can be interactively explored and searched rapidly with ease, whilst numerous graphs provide additional information. The tool can be run locally on any platform. All results are saved in self-contained folders for easy access.

We plan to add the following functionalities to VEMV in the future:

- Dynamic graphs, enabling users to select (in real time) variables to compare, and other features such as zooming in/out.
- Inclusion of a few additional popular lightweight evaluation metrics. The modular design of the software means that adding new metrics is a relatively trivial process.
- Using the saved output from the tool to compare multiple MT systems against one another.

References

Bogdan Babych. 2014. Automated mt evaluation metrics and their limitations. *Tradumàtica*, (12):464–470.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Giménez, Jesús Màrquez, and Lluís. 2010. Asiya: An open toolkit for automatic machine translation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Ondřej 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):63–74.

Nitin Madnani. 2011. ibleu: Interactively debugging & scoring statistical machine translation systems. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, pages 213–214.

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 on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Xingyi Song, Trevor Cohn, and Lucia Specia. 2013. BLEU deconstructed: Designing a better MT evaluation metric. *International Journal of Computational Linguistics and Applications*, (4):29–44.

Milos Stanojevic and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages "414–419". "Association for Computational Linguistics".

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system. In *Proceedings of Language Resources and Evaluation*, pages 2051–2054.