ID: 051

facebook AI Research

VizSeq: A Visual Analysis Toolkit For Text Generation Tasks

Changhan Wang, Anirudh Jain, Danlu Chen, Jiatao Gu

Challenges

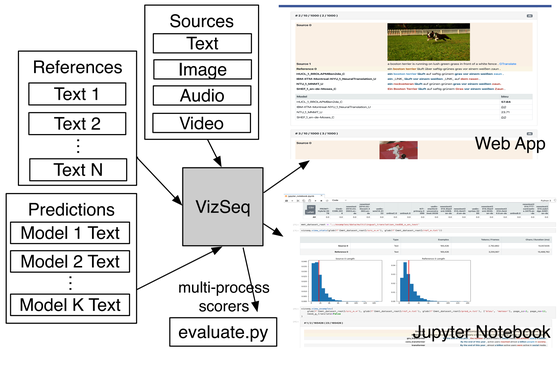
- Wide range of {Text, Image, Audio, Video}-to-text generation tasks: (multimodal) machine translation, summarization, image captioning, speech recognition, video description, etc.
- Automatic metrics are challenged by rich linguistic variations and have gaps towards human judgments

Source	自动评估机器翻译模型是很难的。	BLEU	chrF
Reference	It is difficult to automatically evaluate machine translation models.	100	1.0
Prediction	It is hard to evaluate machine translation models automatically.	34.57	0.74
Prediction	It is easy to automatically evaluate machine translation.	47.50	0.72
Prediction	Machine translation models are difficult to auto-evaluate.	2.16	0.68

- Without looking into examples, abstract scores are limited in model comparison and error analysis
- Inspecting thousands of examples can be a painful process

Our Solution

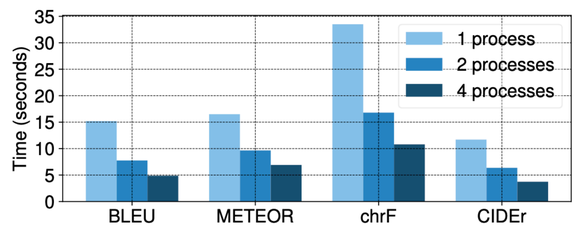
- Highly-Integrated UI with all info (samples, scores, etc.) in one place
- Maximizing productivity with visualization and UI interactions
- Maximizing usability with support of various data sources / formats
- Multi-process computation backend for scalability
- Flexible deployment with Jupyter Notebook and Web App



Main Features

Coverage of Common Metrics, Multi-process Accelerated

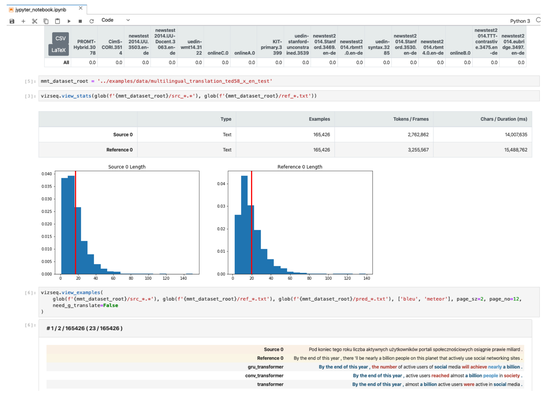
- N-Gram-Based: BLEU, NIST, METEOR, TER, RIBES, chrF, GLEU, ROUGE, CIDEr, WER
- Embedding-Based: LASER, BertScore



Support of Various Data Sources and Formats

- Sources: Python lists/dictionaries, plain text file paths, ZIP file paths
- Formats: txt, jpg, png, wav, flac, mp4, webm, etc.

Jupyter Notebook Interface



```
import vizseq
from glob import glob
src, ref, hypo = glob('src_*.*'), glob('ref_*.txt'), glob('pred_*.txt')
```

```
vizseq.view_examples(src, ref, hypo, ['bleu'], need_g_translate=True)
```

```
vizseq.view_stats(src, ref)
```

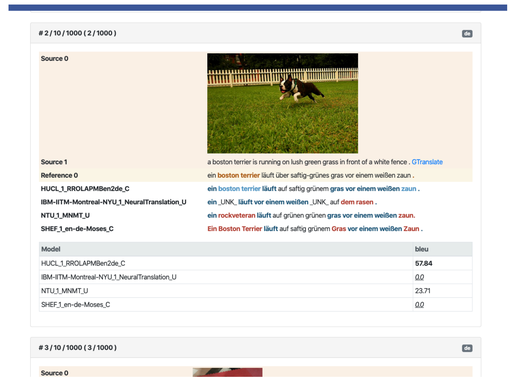
```
vizseq.view_scores(ref, hypo, ['bleu', 'meteor'])
```

```
vizseq.view_n_grams(src)
```

Fairseq Integration

```
from vizseq.ipynb import fairseq_viz as vizseq_fs
```

Web App Interface



```
> python –m vizseq.server --data-root [data_root]
```

Visualization

- hypothesis-reference/reference-source alignments
- Corpus-level statistics
- Highlights of lowest/highest scores

Browsing

- Filtering by n gram keywords, sentence tags, etc.
- Sorting by lengths, sentence scores, etc.
- Grouping by sentence tags
- Google Translate integration

Export

- Figures to PNG or SVG
- Tables to CSV or LaTeX

Check It Out :-)

https://github.com/facebookresearch/vizseq

```
> pip install vizseq
```

References

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(1):63–74.*

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". *NAACL HLT 2019 (2019): 35..*

David Steele and Lucia Specia. (2018). "Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output.". *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 71–75.*