MT-ComparEval: Graphical evaluation interface for Machine Translation development

- Try it now (all WMT 2014–2015 results): http://wmt.ufal.cz
- Install it (and report issues or contribute): https://github.com/choko/MT-ComparEval

- web-based tool for MT developers
- check progress of a system over time or compare several MT systems
- focus on analyzing system differences
- can be integrated into your workflow (import translations from disk/git/...)

Select an "Experiment", e.g. English-Czech WMT15

See BLEU, F-measure,... Select two "Tasks", i.e. system translations

Sentences sorted by diffBLEU

① Select an "Experiment", e.g. English-Czech WMT15
② See BLEU, F-measure,... Select two "Tasks", i.e. system translations
③ Sentences pane with various diffs highlighted
④ Statistics pane (paired bootstrap resampling,...)
⑤ Confirmed n-grams pane
⑥ Unconfirmed n-grams

Click on any n-gram to see all its occurrences

No more excuse for missing examples!

No more excuse for missing significance tests and confidence intervals!

confirmed n-gram = occurs in the reference (light yellow and blue highlight)

improving n-gram = confirmed n-gram occurring in only one of the systems (dark yellow and blue highlight)

worsening n-gram = unconfirmed, occurring in only one of the systems (red highlighting)

diff (LCS underlined in green)

See the sentences with biggest improvement/worsening
See the 1-grams...4-grams with biggest improvement/worsening
Hints for improving the systems