Martin Popel

- **Treex** NLP framework (Treex::Web no progress since last year)

  - SVN → GitHub (https://github.com/ufal/treex)

    pull requests, issues, forks, Travis-CI, smaller codebase

  - Perl scenarios: easy to use, versioned, parameters

    ```
    treex -Len -Ssrc Read::Sentences from=en.txt
    Scen::EN2CS Write::Sentences to=cs.txt
    ```

- **TectoMT** machine translation (PhD thesis on transfer)

  - now for EN↔CS, EN↔ES, EN↔NL, EN↔PT, EN↔EU

  - experiments with Vowpal Wabbit and word embeddings

- **HamleDT 3.0** (42 treebanks), Dan's GAČR **Manyla**

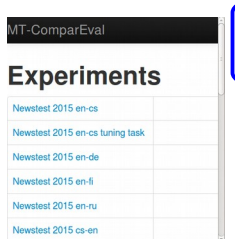- **MT-ComparEval** (with Ondřej Klejch, paper in PBML/MTM)

  https://github.com/choko/MT-ComparEval

MT-ComparEval:
Graphical evaluation interface
for Machine Translation development

Ondřej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, Martin Popel
klejch@ufal.mff.cuni.cz, {eleftherios.avramidis,aljoscha.burchardt}@dfki.de, popel@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
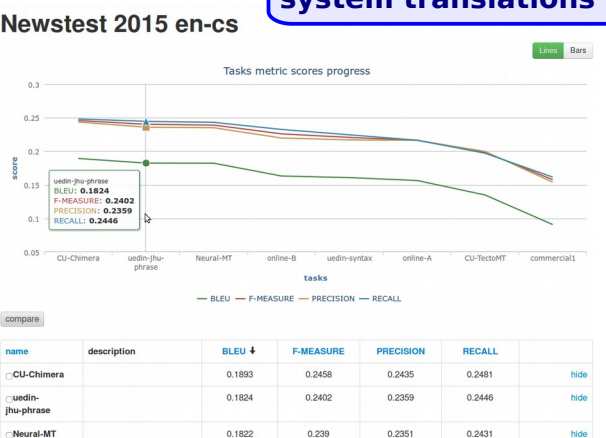German Research Center for Artificial Intelligence (DFKI), Language Technology Lab

● **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/...)

① Select an "Experiment", e.g. English-Czech WMT15

② See BLEU, F-measure,... Select two "Tasks", i.e. system translations

MT-ComparEval — Experiments

Experiments

- Newstest 2015 en-cs
- Newstest 2015 en-cs tuning task
- Newstest 2015 en-de
- Newstest 2015 en-fi
- Newstest 2015 en-ru
- Newstest 2015 cs-en

Newstest 2015 en-cs

Tasks metric scores progress

uedin-jhu-phrase
BLEU: 0.1824
F-MEASURE: 0.2402
PRECISION: 0.2359
RECALL: 0.2446

tasks: CU-Chimera, uedin-jhu-phrase, Neural-MT, online-B, uedin-syntax, online-A, CU-TectoMT, commercial1

BLEU — F-MEASURE — PRECISION — RECALL

name	description	BLEU ↓	F-MEASURE	PRECISION	RECALL	
CU-Chimera		0.1893	0.2458	0.2435	0.2481	hide
uedin-jhu-phrase		0.1824	0.2402	0.2359	0.2446	hide
Neural-MT		0.1822	0.239	0.2351	0.2431	hide

③ Sentences pane with various diffs highlighted

MT-ComparEval   Newstest 2015 en-cs

Neural-MT | CU-Chimera | BLEU-cis

Sentences | Statistics | Confirmed n-grams | Unconfirmed n-grams

Sentences

Options

N-grams higlighting options
- ☑ Highlight confirmed n-grams
- ☑ Highlight improving n-grams
- ☑ Highlight worsening n-grams

Diff higlighting options
- ☐ Show diff with reference
- ◉ Show diff for Neural-MT
- ○ Show diff for CU-Chimera
- ☑ Show diff with each other

Sentences visibility options
- ☑ Show source
- ☑ Show reference
- ☑ Show Neural-MT
- ☑ Show CU-Chimera
- ☑ Show sentence level metrics

Source	The soundtrack is a mix of reggae and rap and the tunes are upbeat .
Reference	Soundtrack je směsicí reggae a rapu a melodie je optimistická .
Neural-MT	The je směs , a rap a melodie jsou optimistická .
CU-Chimera	Soundtrack je směsicí reggae a rapu a melodie jsou optimistické .

	BLEU	BLEU-cis	F-MEASURE	F-MEASURE-cis	PRECISION	PRECISION-cis	RECALL	RECALL-cis
Neural-MT	0.1173	0.1173	0.1864	0.1864	0.1864	0.1864	0.1864	0.1864
CU-Chimera	0.6989	0.6989	0.7025	0.7025	0.7025	0.7025	0.7025	0.7025
Diff	-0.5816	-0.5816	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161

Sentences sorted by diffBLEU

No more excuse for missing examples!

- ● **confirmed n-gram** = occurres 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, occuring 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

④ Statistics pane (paired bootstrap resampling,...)

Sentence-level BLEU-cis differences

Paired Bootstrap Resampling BLEU-cis differences
Neural-MT is worse than uedin-jhu-phrase: p-value=0.367 (not significant)

Neural-MT wins — Neural-MT loses

Bootstrap Resampling BLEU-cis Neural-MT
BLEU-cis lies in 95% confidence interval: [0.1802, 0.1929]

Bootstrap Resampling BLEU-cis uedin-jhu-phrase
BLEU-cis lies in 95% confidence interval: [0.1812, 0.1934]

No more excuse for missing significance tests and confidence intervals!

⑤ Confirmed n-grams pane

⑥ Unconfirmed n-grams

3-gram

Pilot 0.00 wins		Pilot 1.00 wins	
I do to	9	How do I	28
how can I	6	I change the	14

Click on any n-gram to see all its occurrences

Source	Wie öffne ich ein Dokument in Libreoffice ?
Reference	How do I open a document in LibreOffice ?
Pilot 0.00	As I open a document in LibreOffice ?
Pilot 1.00	How do I open a document in Libreoffice ?

Source	Wie verschicke ich eine Datei über Skype ?
Reference	How do I send a file using Skype ?
Pilot 0.00	As I always send a file on Skype ?
Pilot 1.00	How do I send a file above Skype ?

Martin Popel

- **PBML** (next deadline: January 15th 2016) LaTeX volunteer?

- **Technical reports** (2015 deadline: December 1st)

- **Teaching** autumn: Modern Methods in CL I ("Reading group")

  spring: Modern Methods in CL II (for PhD and staff, Deep NN ?)

  Language Data Resources (with ZŽ, me: significance)

- **QTLeap** FP7 project, 2013–2016
  (14 lang. pairs, 8 partners: FCUL, DFKI, CUNI, IICT-BAS, UBER, UPV/EHU, UG, HF)

  TectoMT outperforms Pilot0 (Moses) on the IT domain
  (QTLeapCorpus) for: EN↔CS, EN→PT, EN→ES, NL→EN

  TectoMT available via MT-Monkey web service
  and (more) easily installable