Using MT-ComparEval

Roman Sudarikov, Martin Popel, Ondřej Bojar,
Aljoscha Burchardt, Ondřej Klejch

Charles University in Prague, ÚFAL
DFKI Berlin
Centre for Speech Technology Research, University of Edinburgh

LREC 2016 MT Eval Workshop
Portorož, May 24

What is MT-ComparEval?

- web-based tool for MT developers, who can
- check progress of a system over time
  or compare several MT systems
- focus on analyzing system differences
- integrate MT-ComparEval into their workflow
  (import translations: disk/git/REST-API)

Try it now!

- **http://wmt.ufal.cz**

  all WMT 2014–2016 systems

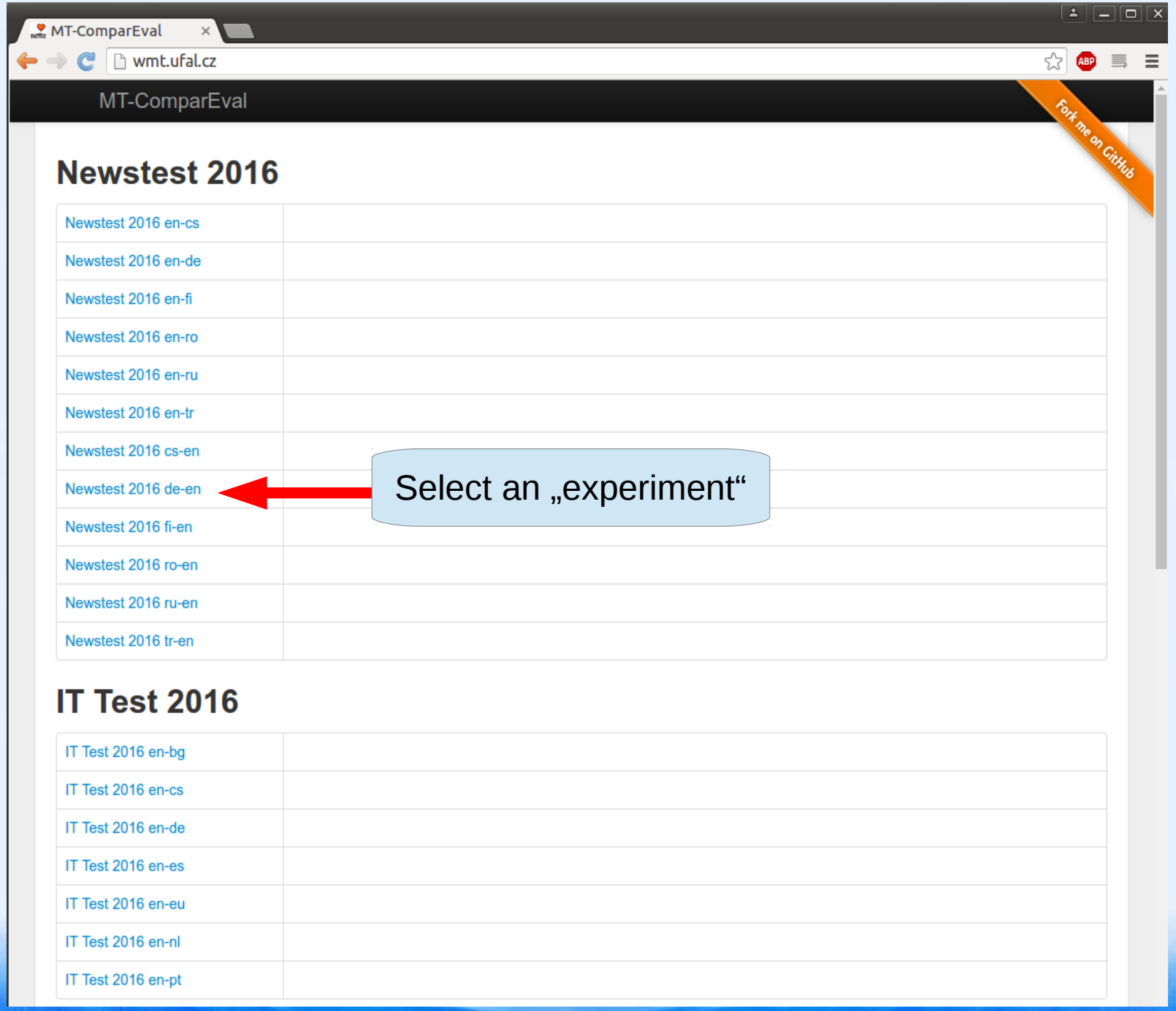
- **http://mt-compareval.ufal.cz**

  upload and inspect your translations

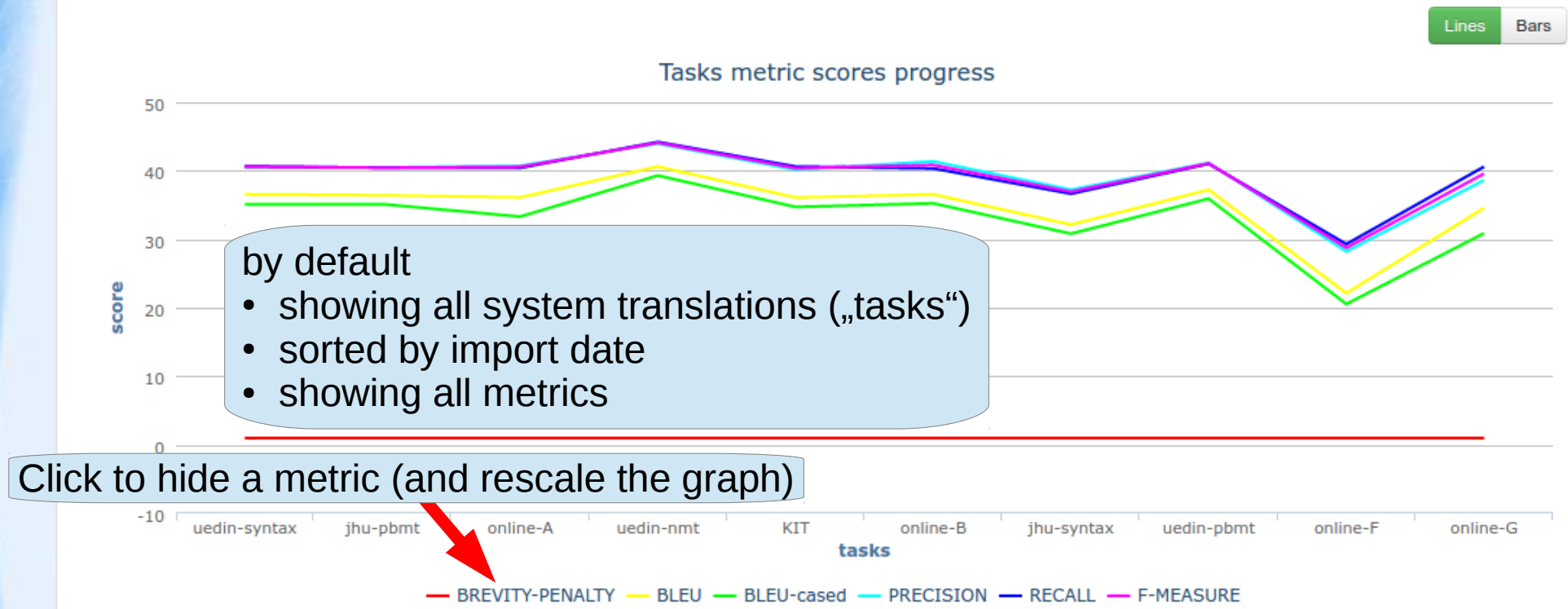
- **https://github.com/choko/MT-ComparEval**

  install it (and report issues or contribute)

wmt.ufal.cz


MT-ComparEval
wmt.ufal.cz
MT-ComparEval
Fork me on GitHub
Newstest 2016
Newstest 2016 en-cs
Newstest 2016 en-de
Newstest 2016 en-fi
Newstest 2016 en-ro
Newstest 2016 en-ru
Newstest 2016 en-tr
Newstest 2016 cs-en
Newstest 2016 de-en
Newstest 2016 fi-en
Newstest 2016 ro-en
Newstest 2016 ru-en
Newstest 2016 tr-en
Select an „experiment“
IT Test 2016
IT Test 2016 en-bg
IT Test 2016 en-cs
IT Test 2016 en-de
IT Test 2016 en-es
IT Test 2016 en-eu
IT Test 2016 en-nl
IT Test 2016 en-pt

Newstest 2016 de-en



by default
- showing all system translations („tasks“)
- sorted by import date
- showing all metrics

Click to hide a metric (and rescale the graph)

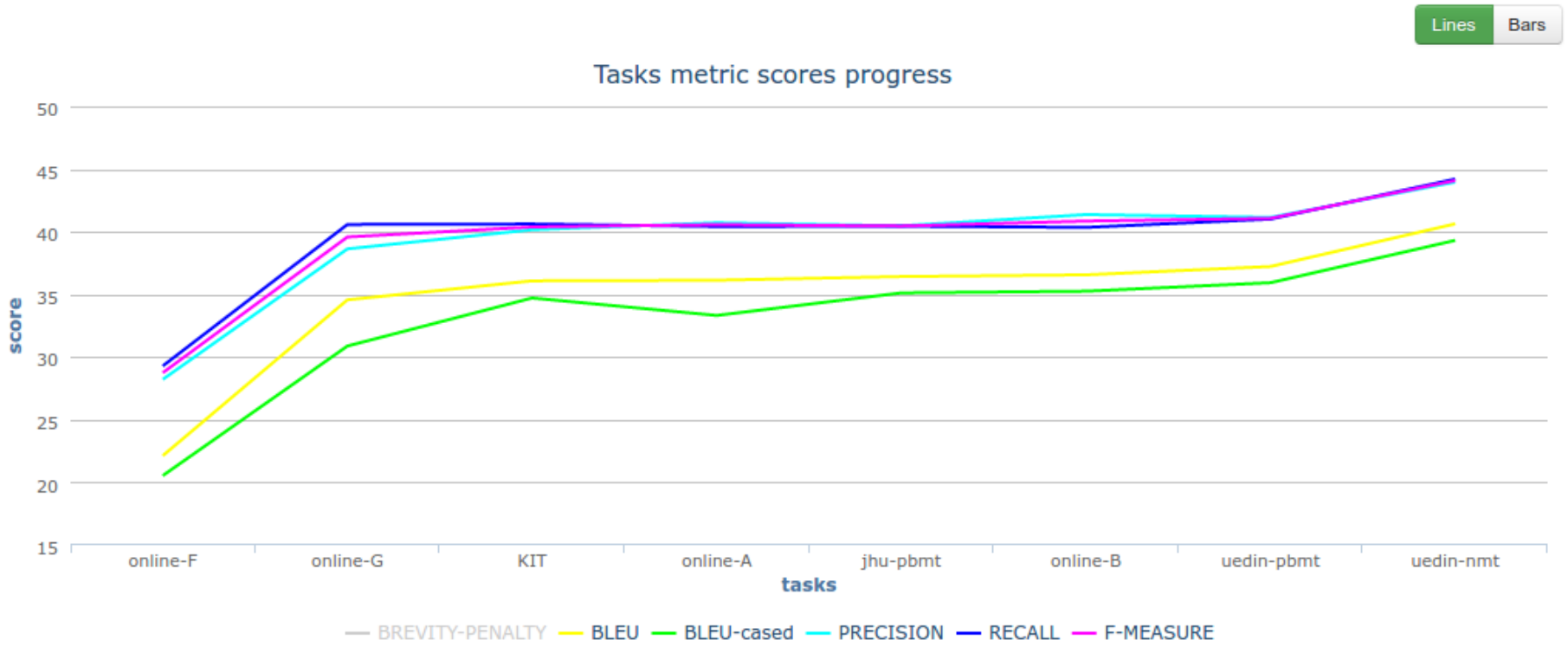
compare

Click to sort by BLEU

Click to hide a system

name	description	BREVITY-PENALTY	BLEU	BLEU-cased	PRECISION	RECALL	F-MEASURE	
uedin-syntax		1	36.57	35.09	40.53	40.67	40.6	hide
jhu-pbmt		1	36.43	35.13	40.47	40.45	40.46	hide
online-A		0.99	36.13	33.32	40.72	40.42	40.57	hide
uedin-nmt		1	40.63	39.32	43.96	44.21	44.08	hide
KIT		1	36.09	34.72	40.16	40.62	40.39	hide
online-B		0.98	36.57	35.26	41.37	40.35	40.85	hide

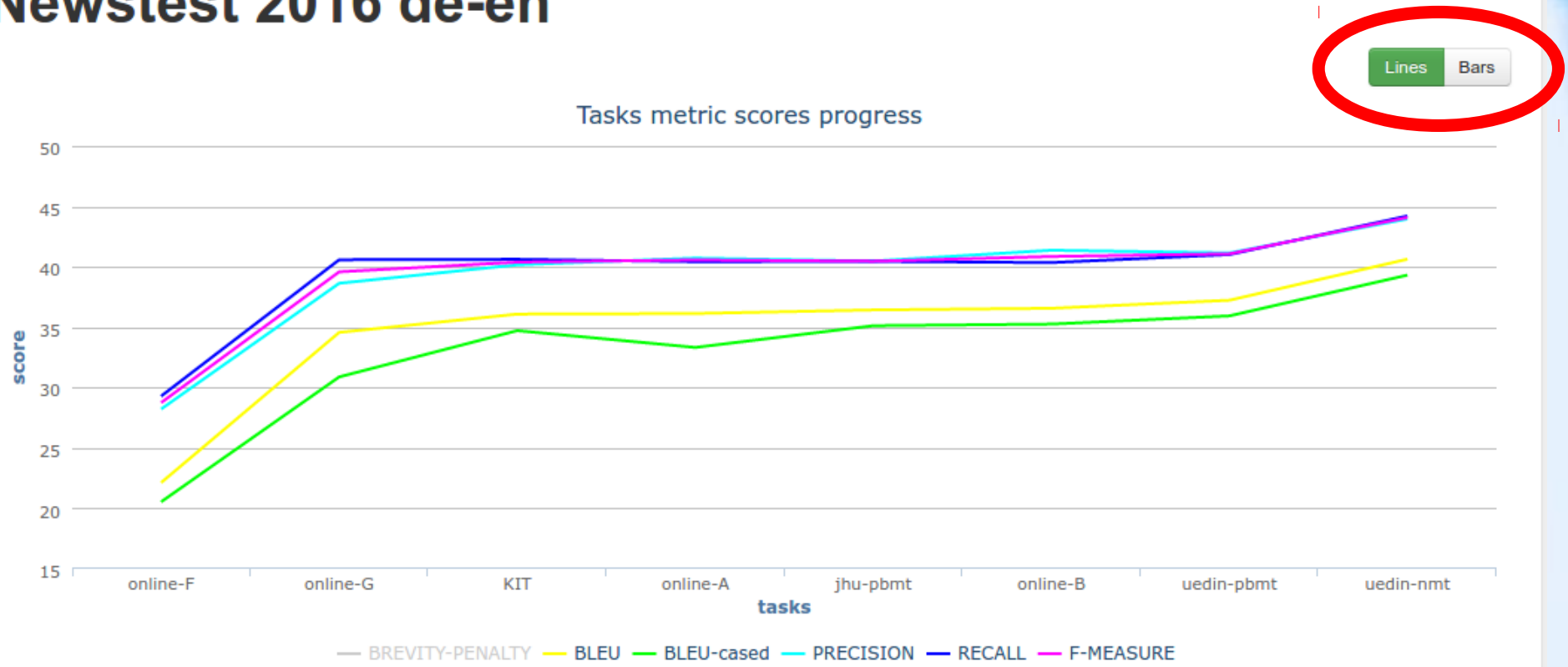
Newstest 2016 de-en



compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

Newstest 2016 de-en

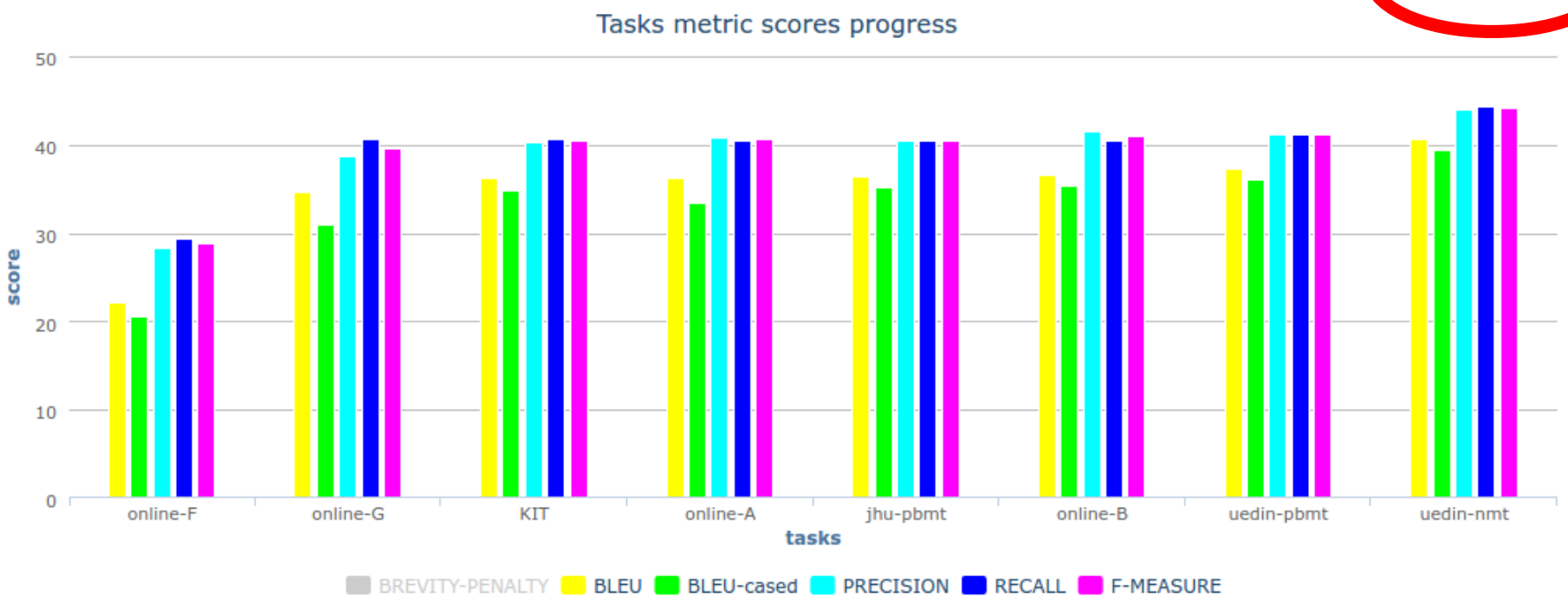


compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

Newstest 2016 de-en

Lines Bars

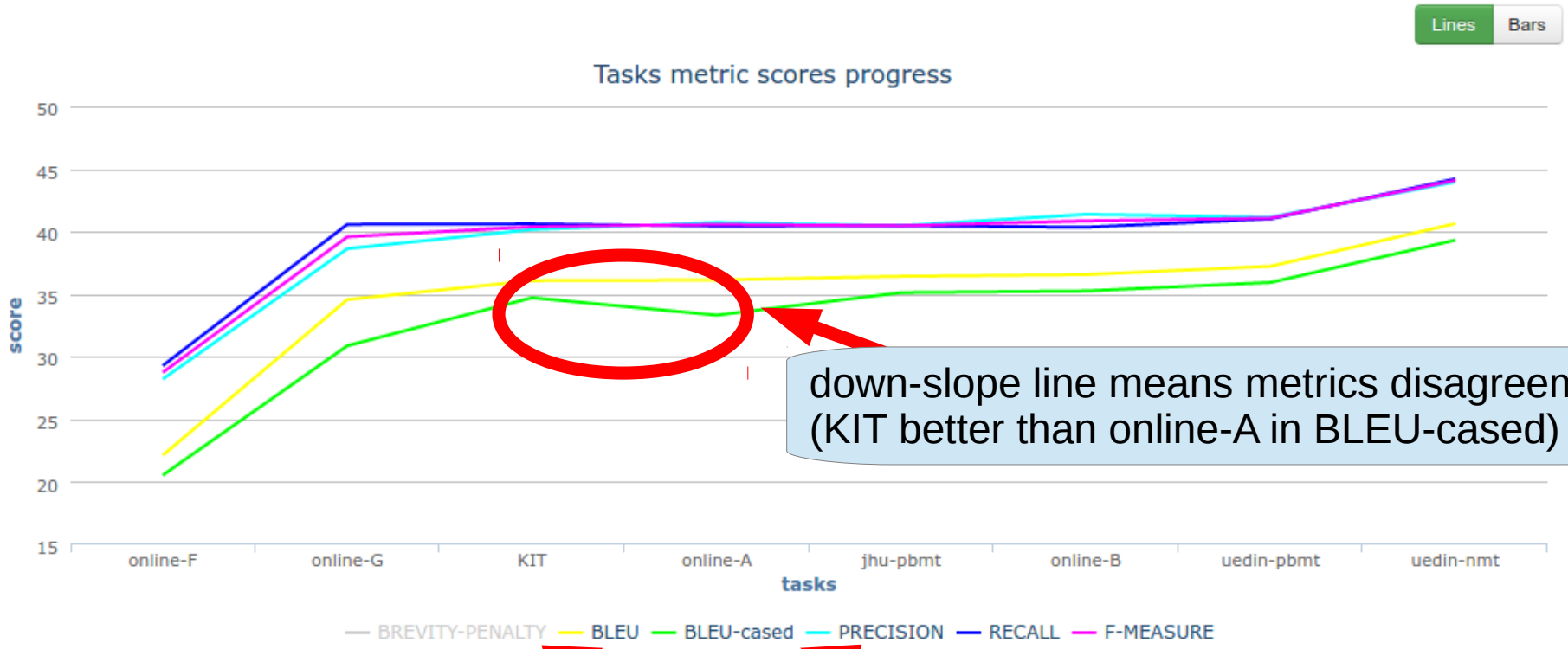


compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

MT-CompareEval

Newstest 2016 de-en



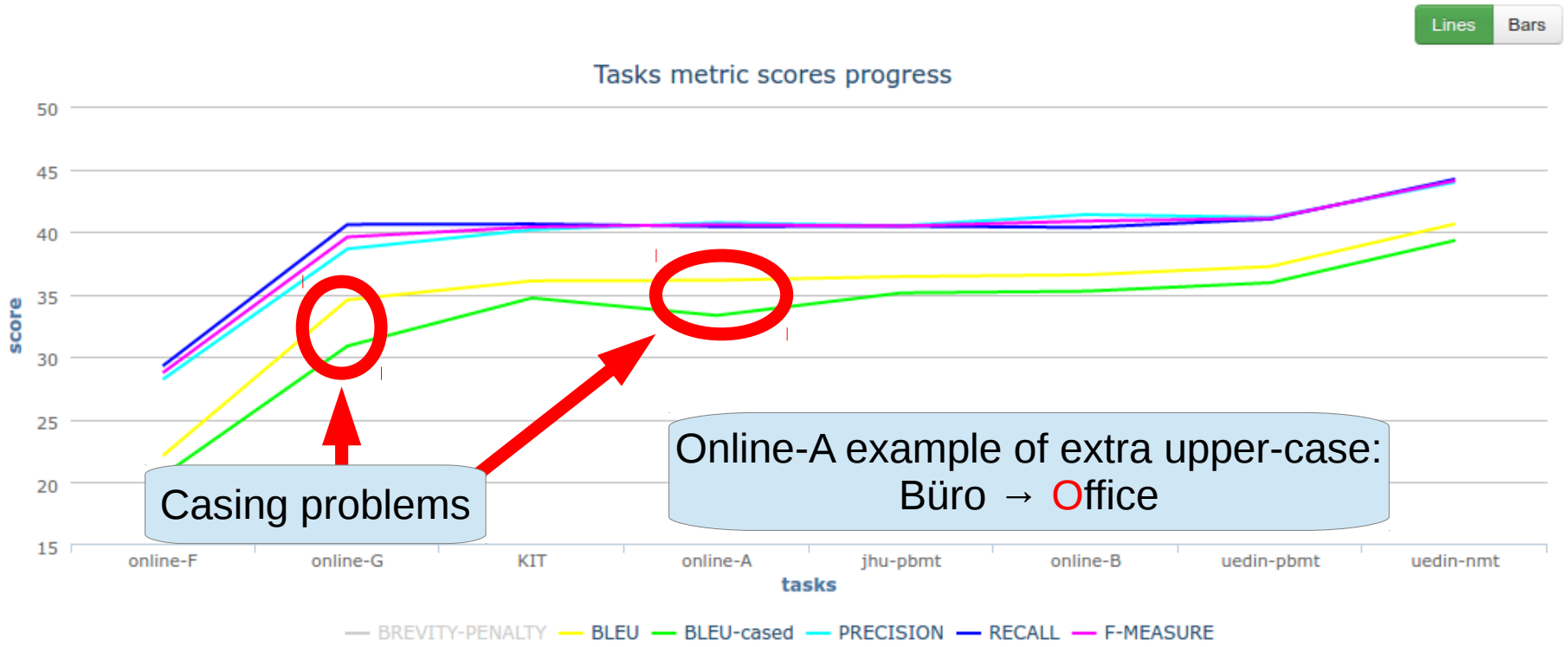
down-slope line means metrics disagreement (KIT better than online-A in BLEU-cased)

BLEU (case insensitive) vs. BLEU-cased

compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

Newstest 2016 de-en



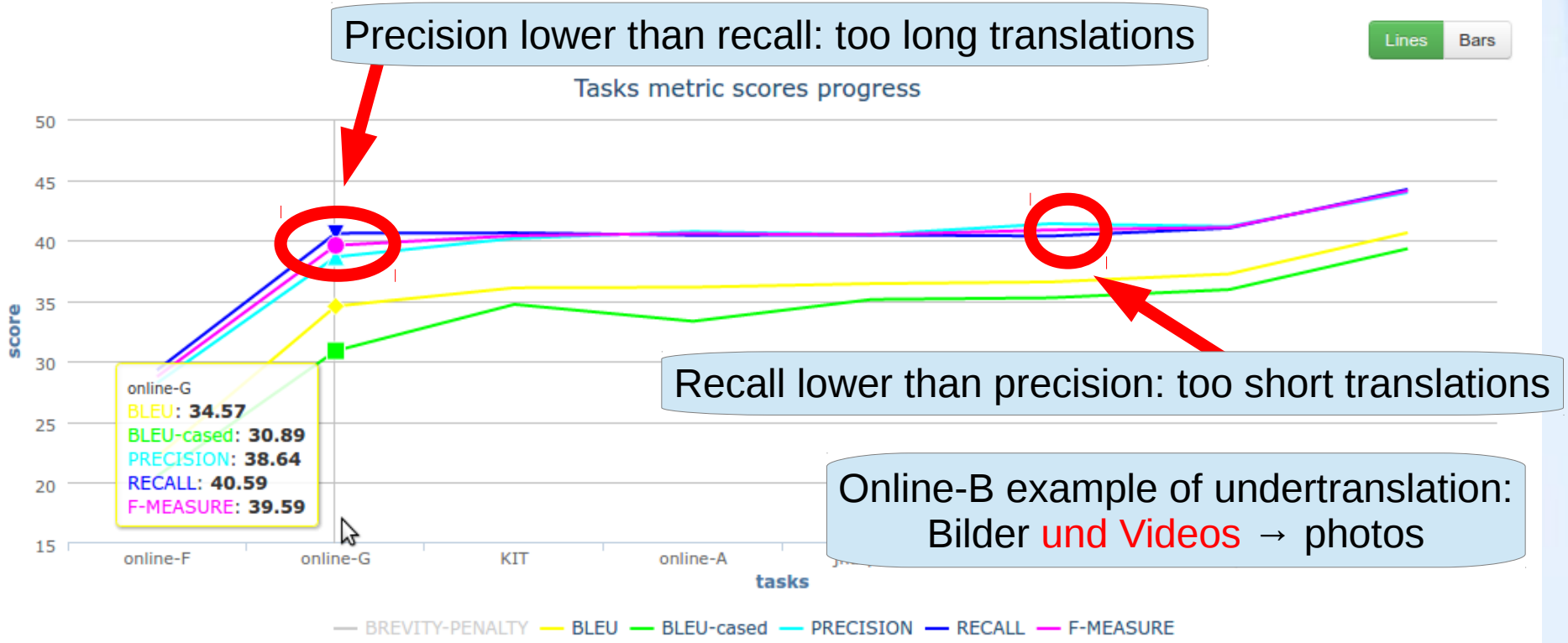
compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

MT-CompareEval

Newstest 2016 de-en

Precision lower than recall: too long translations



Recall lower than precision: too short translations

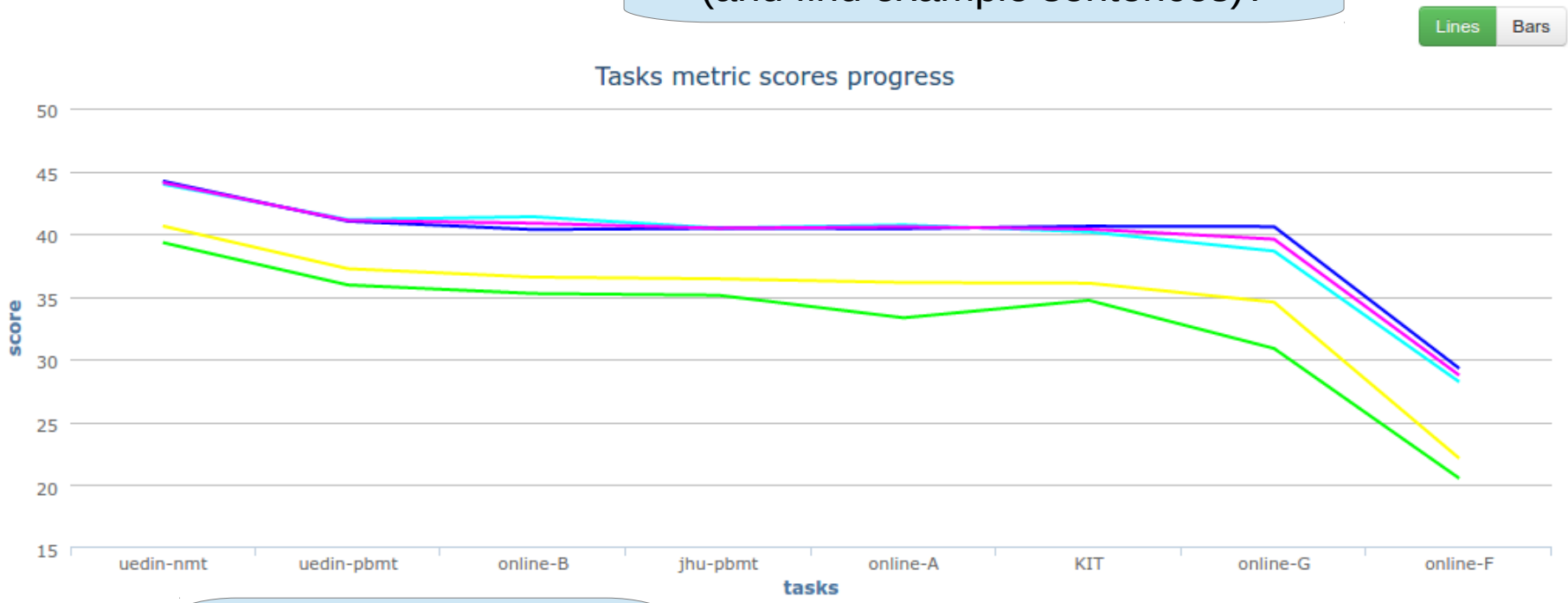
Online-B example of undertranslation:
Bilder und Videos → photos

compare

name	description	BLEU ↑	BLEU-cased	PRECISION	RECALL	F-MEASURE	
online-F	1	22.13	20.55	28.22	29.29	28.75	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide

Newstest 2016 de-en

How to see the differences
(and find example sentences)?



Select the two systems
and click “compare“.

compare

name	description	BLEU ↓	BLEU-cased	PRECISION	RECALL	F-MEASURE	
uedin-nmt	1	40.63	39.32	43.96	44.21	44.08	hide
uedin-pbmt	1	37.23	35.94	41.15	41.02	41.08	hide
online-B	0.98	36.57	35.26	41.37	40.35	40.85	hide
jhu-pbmt	1	36.43	35.13	40.47	40.45	40.46	hide
online-A	0.99	36.13	33.32	40.72	40.42	40.57	hide
KIT	1	36.09	34.72	40.16	40.62	40.39	hide
online-G	1	34.57	30.89	38.64	40.59	39.59	hide

MT-ComparEval Newstest 2016 de-en

online-B | jhu-pbmt | BLEU

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

You can quickly compare with another system

Sentences

Options

N-grams higlighting options

- [x] Highlight confirmed n-grams
- [x] Highlight improving n-grams
- [x] Highlight worsening n-grams

Diff higlighting options

- [ ] Show diff with reference
- (•) Show diff for online-B
- ( ) Show diff for jhu-pbmt
- [ ] Show diff with each other

Sentences visibility options

- [x] Show source
- [x] Show reference
- [x] Show online-B
- [x] Show jhu-pbmt
- [x] Show sentence level metrics

Source	Konkrete Zahlen nannte sie nicht und verwies auf Gespräche am Mittwoch in Berlin .
Reference	She gave no specific figures and referred to talks in Berlin on Wednesday .
online-B	Concrete numbers they did not name , citing talks on Wednesday in Berlin .
jhu-pbmt	She gave no specific figures and referred to talks in Berlin on Wednesday .

	BREVITY-PENALTY	BLEU	BLEU-cased	PRECISION	RECALL	F-MEASURE
online-B	1	8.89	8.89	14.56	14.56	14.56
jhu-pbmt	1	100	100	100	100	100
Diff	0.0000	-91.1100	-91.1100	-85.4400	-85.4400	-85.4400

Source	Das Quartiersbüro ist geöffnet Montag , Mittwoch und Freitag von 9 bis 14 Uhr , Dienstag und Donnerstag von 16 bis 18 Uhr und samstags von 10 Uhr bis 12 Uhr .

MT-ComparEval Newstest 2016 de-en

online-B | jhu-pbmt | BLEU

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

4 panes

the selected metric

Sentences

Options

N-grams higlighting options

- [x] Highlight confirmed n-grams
- [x] Highlight improving n-grams
- [x] Highlight worsening n-grams

Diff higlighting options

- [ ] Show diff with reference
- (•) Show diff for online-B
- ( ) Show diff for jhu-pbmt

Sentences visibility options

- [x] Show source
- [x] Show reference
- [x] Show online-B

Sentences sorted according to the difference in the selected metric.
Top: sentences where JHU-PBMT outperforms Online-B the most.
Very useful also for regression testing (new version vs. baseline).

Reverse ordering
Top: OnlineB wins

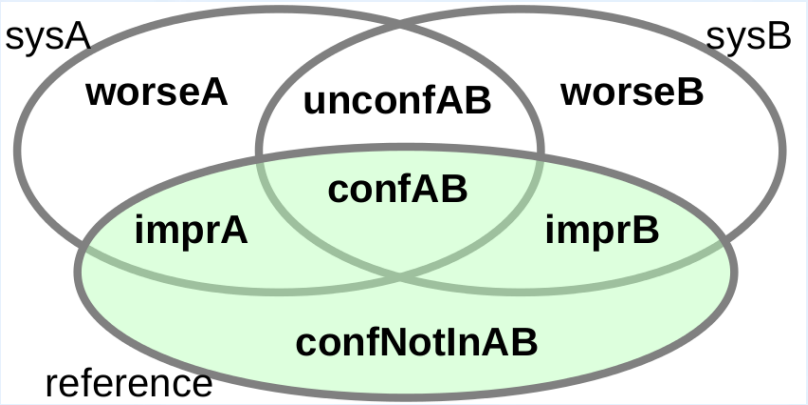
Source	Konkrete Zahlen nannte sie nicht und verwies auf Gespräche am Mittwoch in Berlin .
Reference	She gave no specific figures and referred to talks in Berlin on Wednesday .
online-B	Concrete numbers they did not name , citing talks on Wednesday in Berlin .
jhu-pbmt	She gave no specific figures and referred to talks in Berlin on Wednesday .

	BREVITY-PENALTY	BLEU	BLEU-cased	PRECISION	RECALL	F-MEASURE
online-B	1	8.89	8.89	14.56	14.56	14.56
jhu-pbmt	1	100	100	100	100	100
Diff	0.0000	-91.1100	-91.1100	-85.4400	-85.4400	-85.4400

Source	Das Quartiersbüro ist geöffnet Montag , Mittwoch und Freitag von 9 bis 14 Uhr , Dienstag und Donnerstag von 16 bis 18 Uhr und samstags von 10 Uhr bis 12 Uhr .

Color highlighting

Source	Wie hoch sind die Kosten ?
Reference	What are the costs ?
online-B	How high are the costs ?
jhu-pbmt	What are the costs ?



- **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)

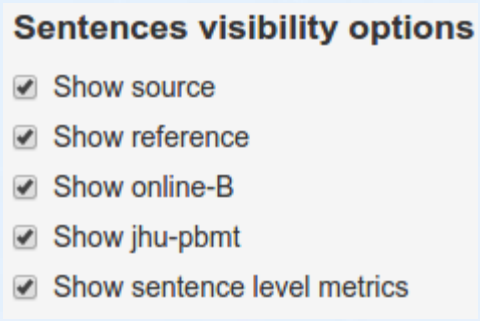
Sentence pane tricks

Source	Sie schicken uns Ihre Bilder und Videos oder eine SMS an 61124.
online-B	Send us your photos or SMS to 61124th
jhu-pbmt	Send us your pictures and videos , or an SMS to 61124.

Source	Fischer ist Spezialist für Herzschrittmacher und Defibrillatoren .
online-B	Fischer is a specialist for pacemakers and defibrillators .
jhu-pbmt	Fischer is a specialist in cardiac pacemakers and defibrillators .

Source	Konrad wollte einen Container am Stephansplatz als Büro aufstellen .
online-B	Konrad wanted to set up a container at Stephansplatz offices .
jhu-pbmt	Konrad wanted to set up a container on Stephansplatz as an office .

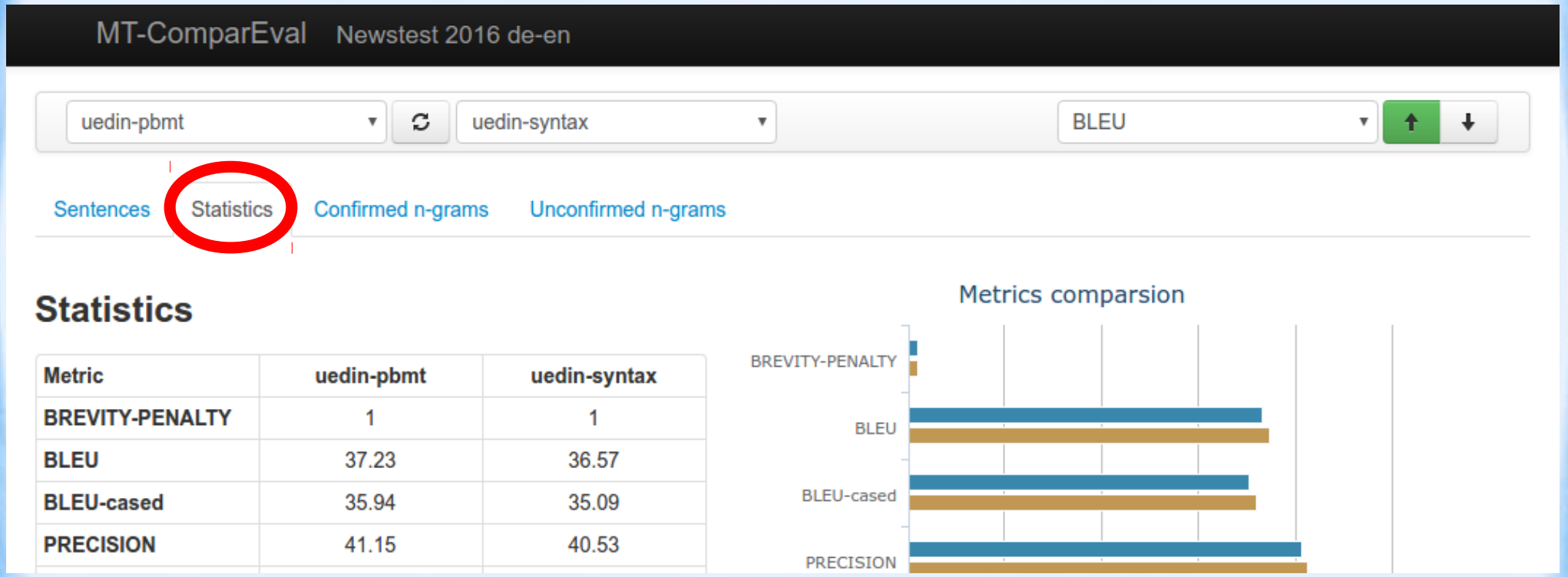
Source	Wie hoch sind die Kosten ?
online-B	How high are the costs ?
jhu-pbmt	What are the costs ?



Looking for nice example sentences?

- Look for short sentences (for slides).
- Search for blue-highlighted content words with no equivalent in the other system's translation.
- Or use Hjerson-omissions metric (off by default).
- Show just the reference and search for blue.
- More sentences are loaded as you scroll down.

Statistics pane



Metric	uedin-pbmt	uedin-syntax
BREVITY-PENALTY	1	1
BLEU	37.23	36.57
BLEU-cased	35.94	35.09
PRECISION	41.15	40.53

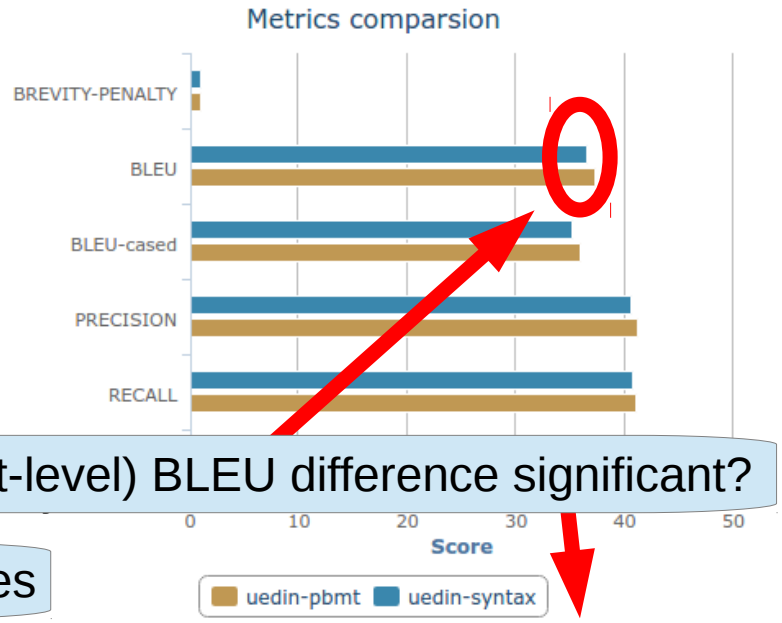
MT-ComparEval Newstest 2016 de-en

uedin-pbmt | uedin-syntax | BLEU

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

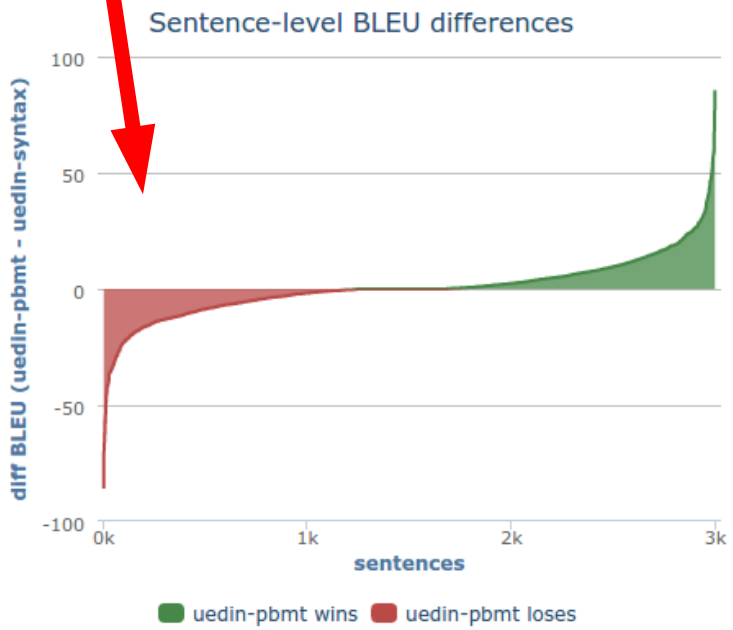
Statistics

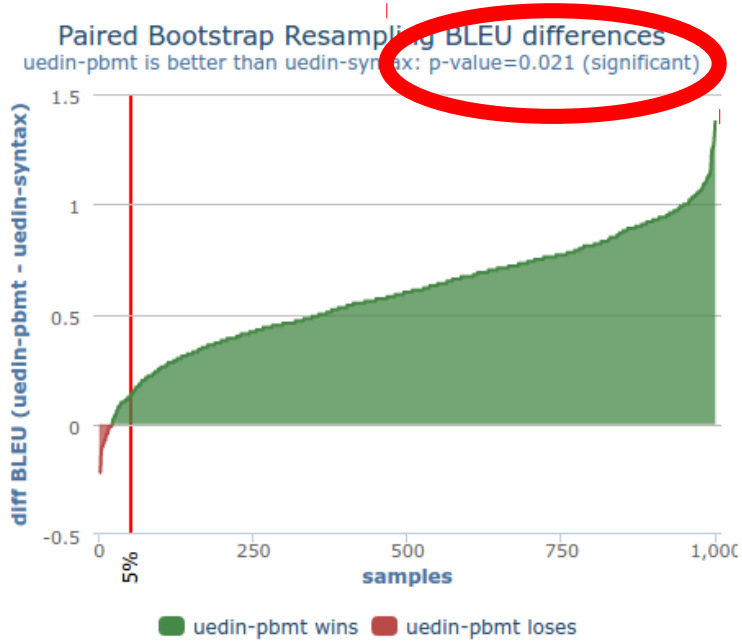
Metric	uedin-pbmt	uedin-syntax
BREVITY-PENALTY	1	1
BLEU	37.23	36.57
BLEU-cased	35.94	35.09
PRECISION	41.15	40.53
RECALL	41.02	40.67
F-MEASURE	41.08	40.6



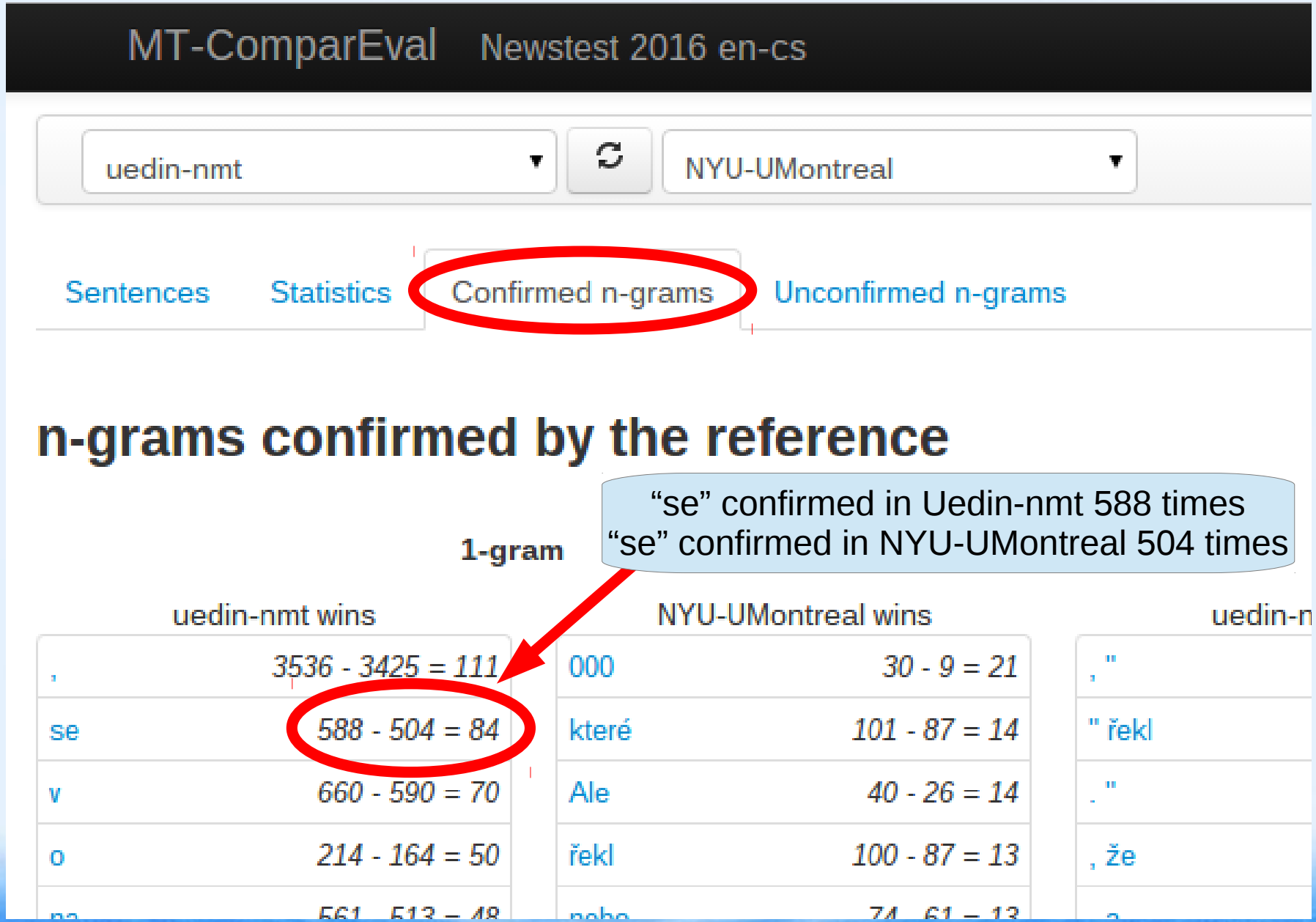
Is the (document-level) BLEU difference significant?

Distribution of sentence-level BLEU differences

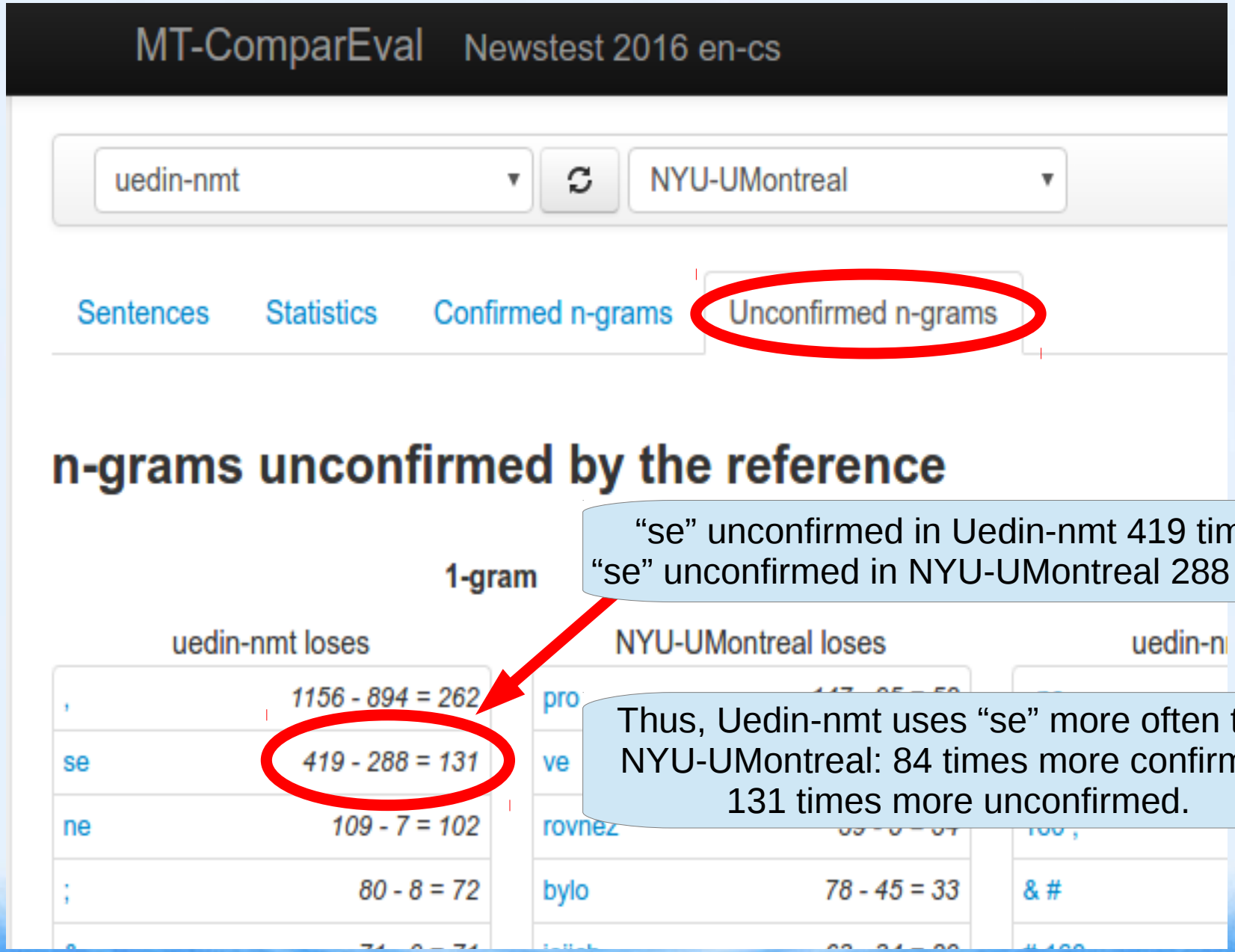




Confirmed n-grams pane



Unconfirmed n-grams pane



MT-ComparEval Newstest 2016 en-cs

uedin-nmt | NYU-UMontreal | BLEU

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

n-grams unconfirmed by the reference

Encoding problems

1-gram

uedin-nmt loses		NYU-UMontreal loses	
,	1156 - 894 = 262	pro	147 - 95 = 52
se	419 - 288 = 131	ve	128 - 91 = 37
ne	109 - 7 = 102	rovněž	39 - 5 = 34
;	80 - 8 = 72	bylo	78 - 45 = 33
&	71 - 0 = 71	jejich	63 - 34 = 29
#	69 - 0 = 69	.	96 - 71 = 25
160	69 - 0 = 69	byla	74 - 51 = 23
o	165 - 98 = 67	být	66 - 43 = 23
v	336 - 284 = 52	řekl	115 - 95 = 20
že	241 - 189 = 52	rok	24 - 4 = 20

2-gram

uedin-nmt loses		NYU-UMontreal loses	
, ne	101 - 2 = 99	" ,	210 - 31 = 179
ne ,	101 - 3 = 98	, řekl	131 - 25 = 106
160 ;	69 - 0 = 69	" .	82 - 48 = 34
& #	69 - 0 = 69	, řekla	20 - 1 = 19
# 160	69 - 0 = 69	, které	91 - 76 = 15
, že	242 - 189 = 53	, protože	22 - 10 = 12
, a	76 - 29 = 47	, říká	23 - 11 = 12
" řekl	42 - 0 = 42	říkal ,	16 - 4 = 12
, který	117 - 95 = 22	říkají ,	15 - 4 = 11
000 &	22 - 0 = 22	předtím ,	12 - 1 = 11

MT-ComparEval Newstest 2016 en-cs

uedin-nmt | NYU-UMontreal | BLEU

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

n-grams unconfirmed by the reference

1-gram

uedin-nmt loses		NYU-UMontreal loses	
,	1156 - 894 = 262	pro	147 - 95 = 52
se	419 - 288 = 131	ve	128 - 91 = 37
ne	109 - 7 = 102	rovněž	39 - 5 = 34
;	80 - 8 = 72	bylo	78 - 45 = 33
&	71 - 0 = 71	jejich	63 - 34 = 29
#	69 - 0 = 69	.	96 - 71 = 25
160	69 - 0 = 69	byla	74 - 51 = 23
o	165 - 98 = 67	být	66 - 43 = 23
v	336 - 284 = 52	řekl	115 - 95 = 20
že	241 - 189 = 52	rok	24 - 4 = 20

2-gram

uedin-nmt loses		NYU-UMontreal loses	
, ne	101 - 2 = 99	" ,	210 - 31 = 179
ne ,	101 - 3 = 98	, řekl	131 - 25 = 106
160 ;	69 - 0 = 69	" .	82 - 48 = 34
& #			20 - 1 = 19
# 160			91 - 76 = 15
, že			22 - 10 = 12
, a	76 - 29 = 47	, říká	23 - 11 = 12
" řekl	42 - 0 = 42	říkal ,	16 - 4 = 12
, který	117 - 95 = 22	říkají ,	15 - 4 = 11
000 &	22 - 0 = 22	předtím ,	12 - 1 = 11

98 vs. 0
What is this?
Let's click on the ngram.

3-gram

uedin-nmt loses		NYU-UMontreal loses	
, ne ,	100 - 1 = 99	" , řekl	100 - 1 = 99
ne , ne	98 - 0 = 98	" , řekla	13 - 0 = 13
& # 160	69 - 0 = 69	říkal , že	16 - 4 = 12
# 160 ;	69 - 0 = 69	říkají , že	14 - 3 = 11
, " řekl	42 - 0 = 42	v České republice	12 - 1 = 11
000 & #	22 - 0 = 22	" , říká	11 - 0 = 11
, že se	49 - 28 = 21	říkala , že	12 - 2 = 10
, a to	25 - 4 = 21	, která byla	13 - 4 = 9
; 000 &	21 - 0 = 21	" , prohlásil	9 - 0 = 9
160 ; 000	21 - 0 = 21	říká , že	22 - 14 = 8

4-gram

uedin-nmt loses		NYU-UMontreal loses	
, ne , ne	98 - 0 = 98	" , řekl .	7 - 0 = 7
ne , ne ,	98 - 0 = 98	, že je to	18 - 14 = 4
& # 160 ;	69 - 0 = 69	, řekl Navrátil .	4 - 0 = 4
000 & # 160	22 - 0 = 22	" , řekl Wenger	4 - 0 = 4
; 000 & #	21 - 0 = 21	, řekl Wenger .	4 - 0 = 4
# 160 ; 000	21 - 0 = 21	" , řekl Navrátil	4 - 0 = 4
160 ; 000 &	21 - 0 = 21	, že to byla	4 - 0 = 4
; & # 160	18 - 0 = 18	v Československé armády by	4 - 0 = 4
# 160 ; &	18 - 0 = 18	" , řekla .	4 - 0 = 4
160 ; & #	18 - 0 = 18		

We see the Sentence pane with a filter for sentences containing n-gram “ne, ne,” which is unconfirmed only in Uedin-nmt (ie. worsening).

You are displaying sentences with worsening n-gram **, ne , ne**. Show all senteces

Source	A two - out walk to right fielder J . D . Martinez brought up Victor Martinez , who singled up the middle for the first run of the game .
Reference	Dva odchody pro pravého polaře J . D . Martineze vynesly Victora Martineze , který jako první oběhl všechny mety .
uedin-nmt	Ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne , ne ,
NYU-UMontreal	Martinez vychoval Victora Martineze , který zpíval uprostřed prvního běhu hry .

There is just one such translation. It contains 100 tokens “ne” (no).

Uedin-nmt is the overall winner of WMT16 (including en→cs).

Apparently, there are still ways how to improve it.