Results of the WMT16 Metrics Shared Task

Ondřej Bojar
Yvette Graham
Amir Kamran
Miloš Stanojević

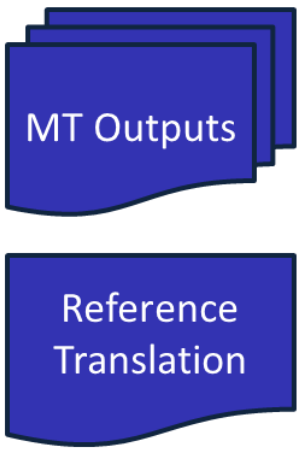
WMT16, Aug 11, 2016

Overview

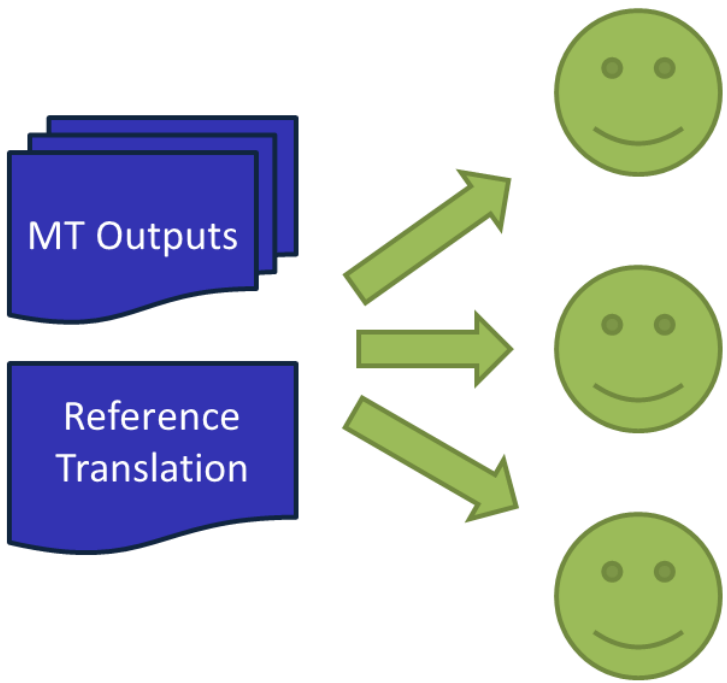
- Summary of Metrics Task.
- Updates to Metric Task in 2016.
- Results

Metrics Task in a Nutshell

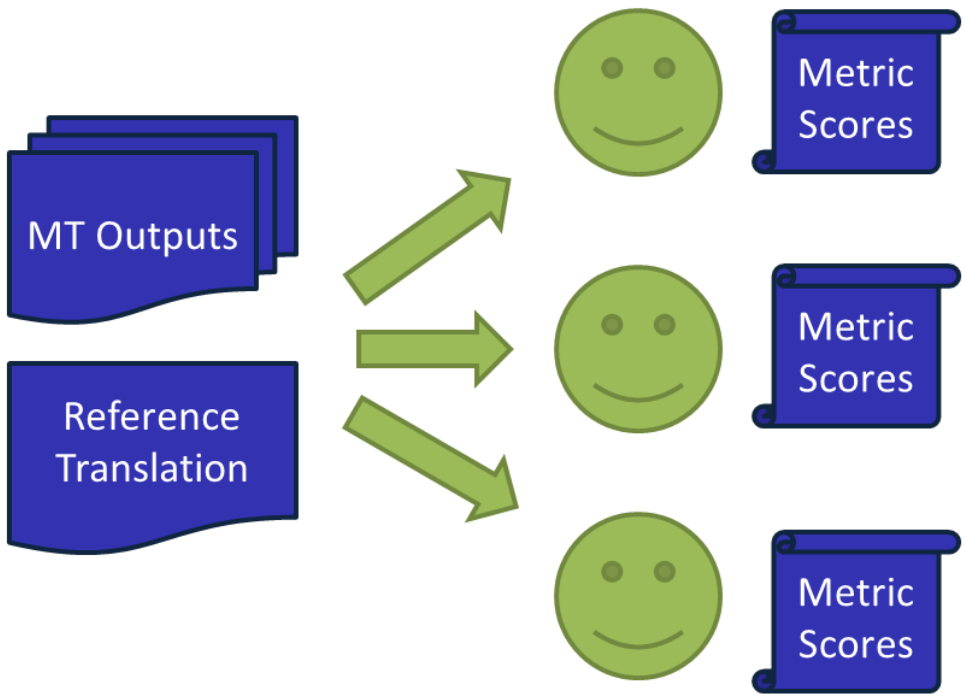
Metrics Task in a Nutshell



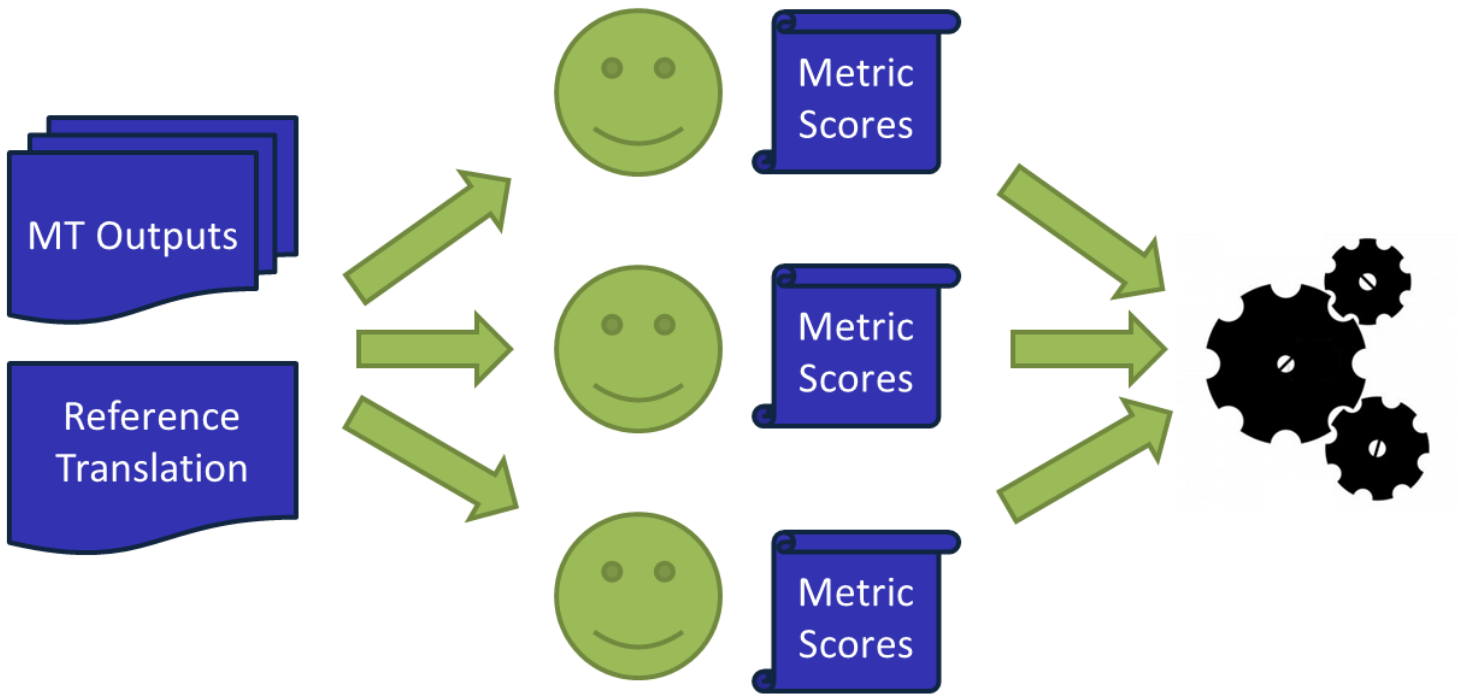
Metrics Task in a Nutshell



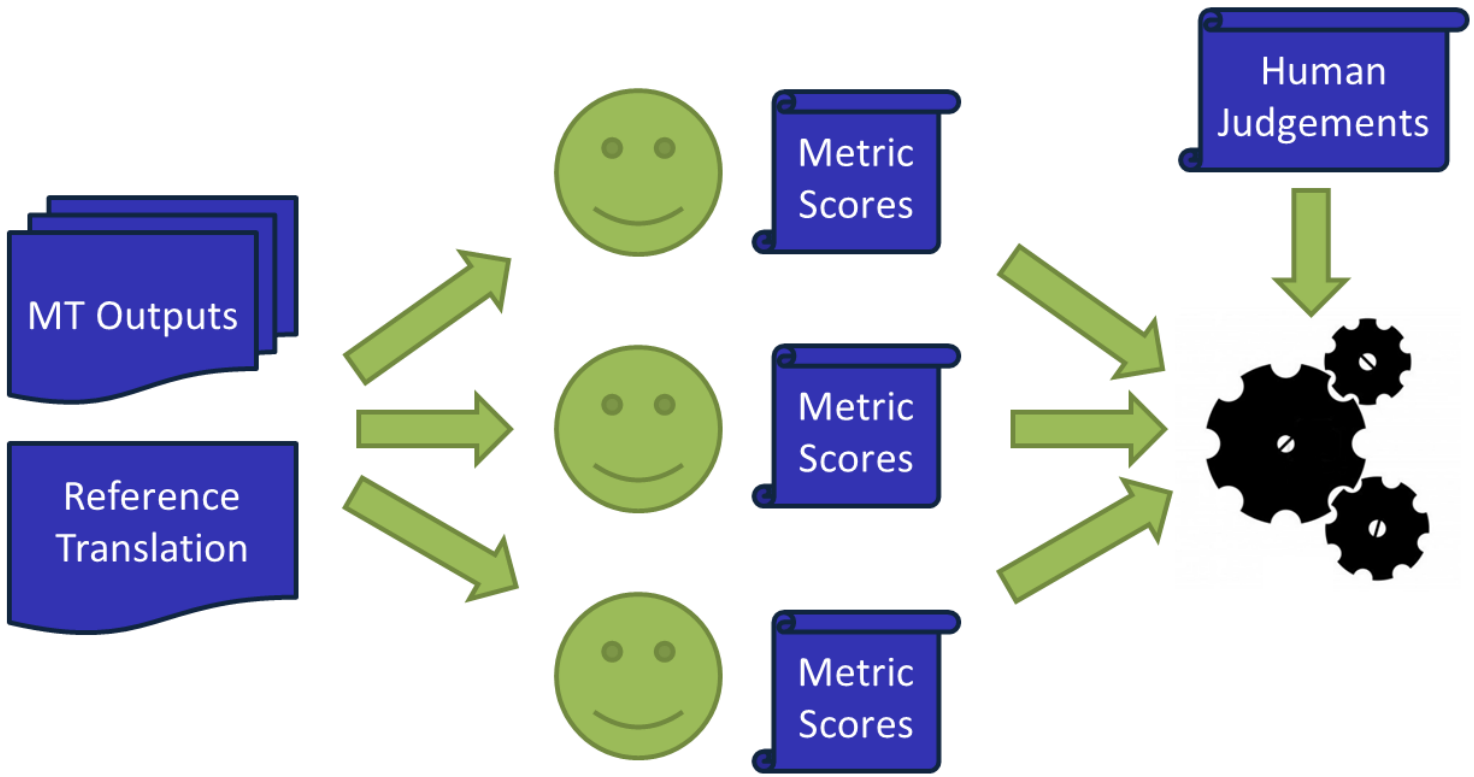
Metrics Task in a Nutshell



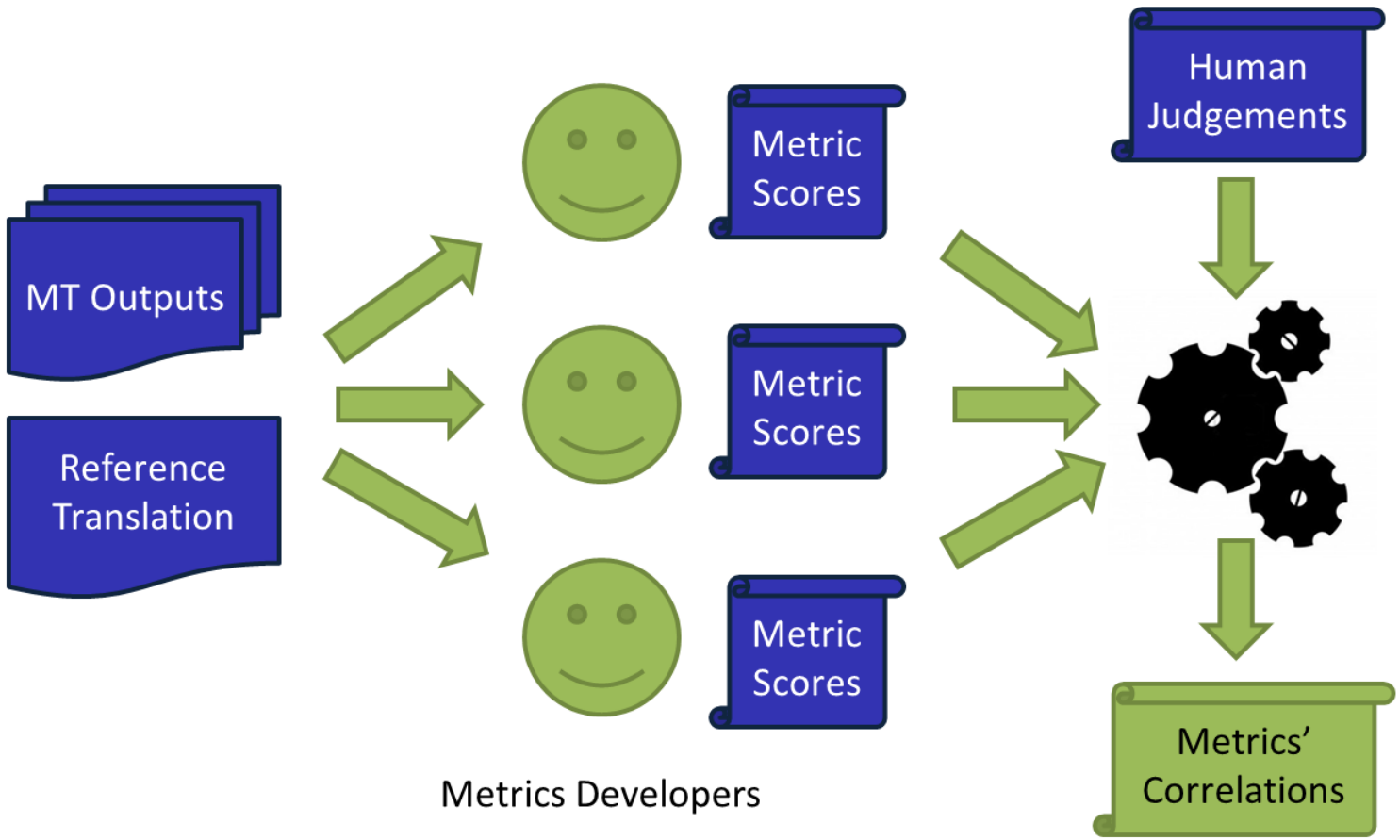
Metrics Task in a Nutshell



Metrics Task in a Nutshell

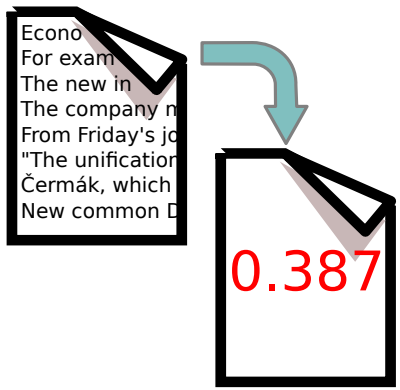


Metrics Task in a Nutshell



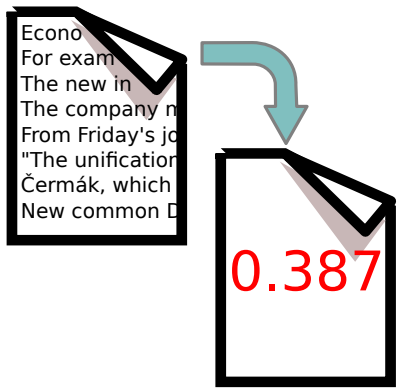
System- and Segment-Level Evaluation

- System Level
  - Participants compute one score for the whole test set, as translated by each of the systems

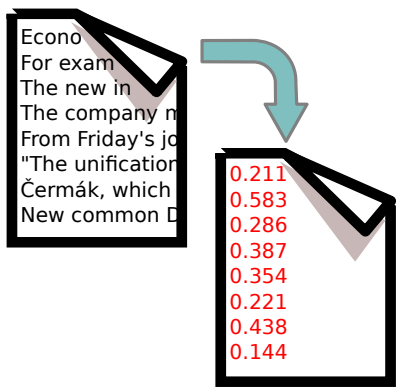


System- and Segment-Level Evaluation

- System Level
  - Participants compute one score for the whole test set, as translated by each of the systems



- Segment Level
  - Participants compute one score for each sentence of each system's translation



Nine Years of Metrics Task

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Participating Teams	-	6	8	14	9	8	12	12	11	**9**
Evaluated Metrics	11	16	38	26	21	12	16	23	46	**16**
Baseline Metrics					2		5	6	7	**9**
System-level										
Spearman Rank Corr.	•	•	•	•	•	•	•	○		
Pearson Corr. Coeff.							○	•	•	•
Segment-level										
Ratio of Concordant Pairs		•								
Kendall's τ				•	•	•	$*$	$\star$	$\star$	$\star$
Pearson Corr. Coeff.										•

• main and ○ secondary score reported for the system-level evaluation.

•, $*$ and $\star$ are slightly different variants regarding ties.

- Stable number of participating teams.
- A growing set of "baseline metrics".
- Stable but gradually improving evaluation methods.

Updates to Metrics Task in 2016

- More Domains
  - News, IT, Medical.
- Two Golden Truths in News Task
  - Relative Ranking, Direct Assessment.
- Third golden truth in Medical Domain.
- Confidence for Sys-level Computed Differently.
  - Participants needed to score 10K systems.
- More languages (18 pairs):
  - Basque, Bulgarian, Czech, Dutch, Finnish, German, Polish, Portuguese, Romanian, Russian, Spanish, and Turkish
  - Paired with English in one or both directions.

Metrics Task Madness

Track	Test set	News Task	Tuning Task	IT Task	HimL Year 1	Hybrid	into-English: cs	de	ro	fi	ru	tr	English into: cs	de	ro	fi	ru	tr	bg	es	eu	nl	pl	pt
		Systems					into-English						English into											
RRsysNews	newstest2016	✓	✓			✓	●	●	●	●	●	●	●	●	●	●	●	●						
RRsysIT	it-test2016			✓		✓							●	●					●	●	●	●		●
DAsysNews	newstest2016	✓	✓			✓	●	●	●	●	●	●	·	·	·	·	●	·						
RRsegNews	newstest2016	✓	✓				●	●	●	●	●	●	●	●	●	●	●	●						
DAsegNews	newstest2016	✓					●	●	●	●	●	●					●							
HUMEseg	himl2015				✓								●	●	●								●	

"✓": sets of underlying MT systems

"●": language pairs covered in the evaluation

"·" language pairs planned but abandoned

Metrics Task Madness

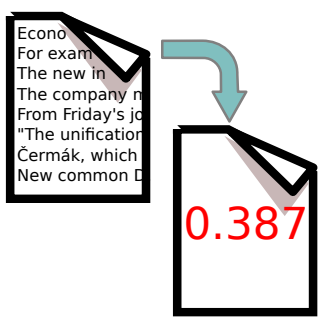
Track	Test set	News Task	Tuning Task	IT Task	HimL Year 1	Hybrid	cs	de	ro	fi	ru	tr	cs	de	ro	fi	ru	tr	bg	es	eu	nl	pl	pt
		Systems					into-English						English into											
RRsysNews	newstest2016	✓	✓			✓	●	●	●	●	●	●	●	●	●	●	●	●						
RRsysIT	it-test2016			✓		✓							●	●					●	●	●	●		●
DAsysNews	newstest2016	✓	✓			✓	●	●	●	●	●	●	·	·	·	·	●	·						
RRsegNews	newstest2016	✓	✓				●	●	●	●	●	●	●	●	●	●	●	●						
DAsegNews	newstest2016	✓					●	●	●	●	●	●					●							
HUMEseg	himl2015				✓								●	●	●								●	

"✓": sets of underlying MT systems

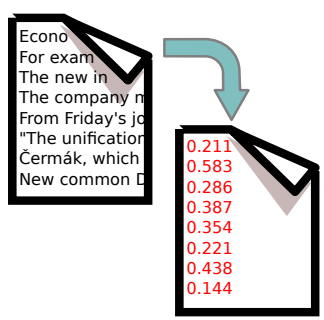
"●": language pairs covered in the evaluation

"·" language pairs planned but abandoned

For participants, this was cut down to the standard:



Sys-level and seg-level scoring.



Metrics Task Domains

- WMT16 News Task
  - Systems and language pairs from the main translation task.
  - Truth: Primarily RR, DA into English and Russian.
- WMT16 IT Task
  - IT domain.
  - Only out of English.
  - Interesting target languages: (Czech, German,) Bulgarian, Spanish, Basque, Dutch, Portuguese.
  - Truth: Only RR
- HimL Medical Texts
  - Just one system per target language.
  - (So only seg-level evaluation.)
  - Truth: A new semantics-based metric.

Golden Truths

- Relative Ranking (RR)
  - 5-way relative comparison.
    - Interpreted as 10 pairwise comparison.
    - Identical outputs deduplicated.
    - Finally converted to a score using TrueSkill.
- Direct Assessment (DA)
  - Absolute adequacy judgement over individual sentences.
    - Judgements from each worker standardized.
    - Multiple judgements of a candidate averaged.
    - Finally averaged over all sentences of a system.
  - Fluency optionally to resolve ties.
  - Provided by Turkers (only English and Russian).
    - Planned but not done with Researchers.
- HUME
  - A composite score of manual judgements of meaning preservation.
  - Used only in the "medical" track.

Effects of DA vs. RR for Metrics Task

Benefits:

- More principled golden truth.
- Possibly more reliable, *assuming enough judgements*.

Negative aspects:

- Sampling for sys-level and seg-level is different.
- Perhaps impossible for seg-level out of English:
  - Too few Turker annotations.
  - ~~Too few researchers.~~ (Repeated judgements work as well.)

This year, only English and Russian news systems have DA judgements.

Participants

Metric	Participant
BEER	ILLC – UvA (Stanojević and Sima'an, 2015)
CharacTer	RWTH Aachen University (Wang et al., 2016)
chrF1,2,3	Humboldt University of Berlin (Popović, 2016)
wordF1,2,3	Humboldt University of Berlin (Popović, 2016)
DepCheck	Charles University, no corresponding paper
DPMFcomb-	Chinese Academy of Sciences
-without-RED	and Dublin City University (Yu et al., 2015)
MPEDA	Jiangxi Normal University (Zhang et al., 2016)
UoW.ReVal	University of Wolverhampton (Gupta et al., 2015)
upf-cobalt	Universitat Pompeu Fabra (Fomicheva et al., 2016)
CobaltF	Universitat Pompeu Fabra (Fomicheva et al., 2016)
MetricsF	Universitat Pompeu Fabra (Fomicheva et al., 2016)
DTED	University of St Andrews, (McCaffery and Nederhof, 2016)

Standard Presentation of the Results

	cs-en		de-en		fi-en		ro-en		ru-en		tr-en	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	6	6	10	10	9	9	7	7	10	10	8	8
MPEDA	**.996**	**.993**	**.956**	.937	**.967**	**.976**	**.938**	**.932**	**.986**	.929	**.972**	**.982**
UoW.ReVal	**.993**	**.986**	**.949**	**.985**	**.958**	**.970**	**.919**	**.957**	**.990**	**.976**	**.977**	**.958**
BEER	**.996**	**.990**	.949	.879	**.964**	**.972**	**.908**	.852	**.986**	.901	**.981**	**.982**
chrF1	.993	.986	.934	.868	**.974**	**.980**	.903	.865	**.984**	.898	**.973**	.961
chrF2	**.992**	.989	.952	.893	**.957**	**.967**	.913	.886	**.985**	.918	.937	.933
chrF3	.991	.989	**.958**	.902	.946	.958	.915	**.892**	**.981**	.923	.918	.917
CharacTer	**.997**	**.995**	**.985**	.929	.921	.927	**.970**	**.883**	.955	**.930**	.799	.827
mtevalNIST	.988	.978	.887	.801	.924	.929	.834	.807	.966	.854	**.952**	**.938**
mtevalBLEU	**.992**	**.989**	.905	.808	.858	.864	.899	.840	.962	.837	.899	.895
mosesCDER	**.995**	**.988**	.927	.827	.846	.860	.925	.800	.968	.855	.836	.826
mosesTER	**.983**	**.969**	.926	.834	.852	.846	.900	.793	.962	.847	.805	.788
wordF2	**.991**	**.985**	.897	.786	.790	.806	.905	.815	.955	.831	.807	.787
wordF3	**.991**	**.985**	.898	.787	.786	.803	.909	.818	.955	.833	.803	.786
wordF1	**.992**	**.984**	.894	.780	.796	.808	.890	.804	.954	.825	.806	.776
mosesPER	.981	.970	.843	.730	.770	.767	.791	.748	**.974**	.887	**.947**	.940
mosesBLEU	**.991**	**.983**	.880	.757	.752	.759	.878	.793	.950	.817	.765	.739
mosesWER	**.982**	**.967**	.926	.822	.773	.768	.895	.762	.958	.837	.680	.651
	newstest2016											

- Bold in RR indicates "official winners".
- Some setups fairly non-discerning, here e.g. csen:
  - All but chrF1, chrF3, mtevalNIST and mosesPER tie at top.

News RR Winners Across Languages

Metric	# Wins	Language Pairs
BEER	11	csen, encs, ende, enfi, enro, enru, entr, fien, roen, ruen, tren
UoW.ReVal	6	csen, deen, fien, roen, ruen, tren
chrF2	6	csen, encs, enro, entr, fien, ruen
chrF1	5	encs, enro, fien, ruen, tren
chrF3	4	deen, enfi, entr, ruen
mosesCDER	4	csen, enfi, enru, entr
CharacTer	3	csen, deen, roen
mosesBLEU	3	csen, encs, enfi
mosesPER	3	enro, ruen, tren
mtevalBLEU	3	csen, encs, enro
wordF1	3	csen, encs, enro
wordF2	3	csen, encs, enro
mosesTER	2	csen, encs
mtevalNIST	2	encs, tren
wordF3	2	csen, entr
mosesWER	1	csen

Graphical Presentation of Significant Wins

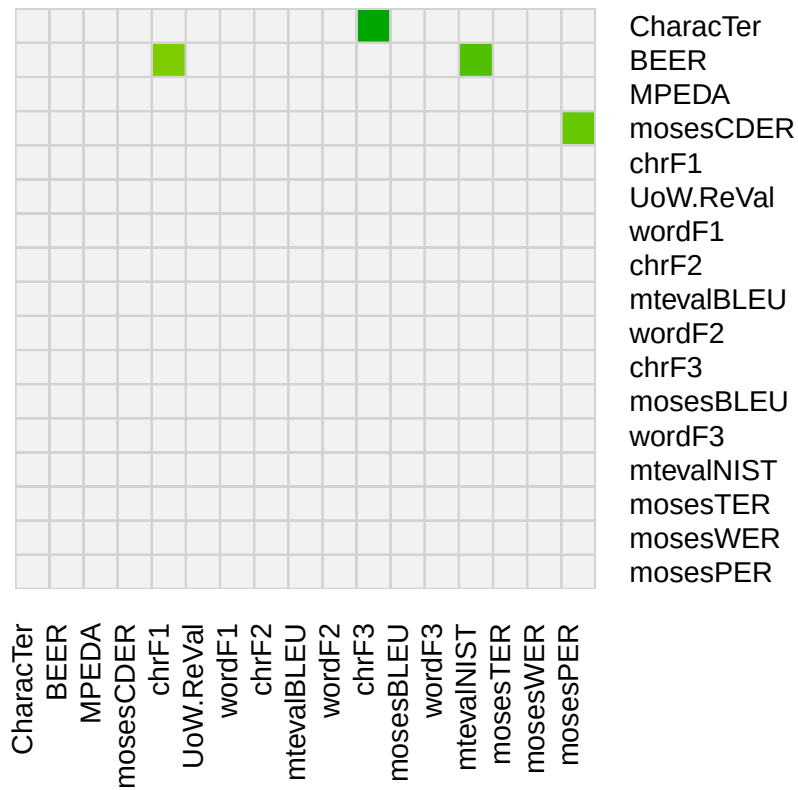
- Williams (1959) test of significant improvement in Pearson correlation.
  - Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

So for Czech-English RR, we have:

Graphical Presentation of Significant Wins

- Williams (1959) test of significant improvement in Pearson correlation.
  - Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

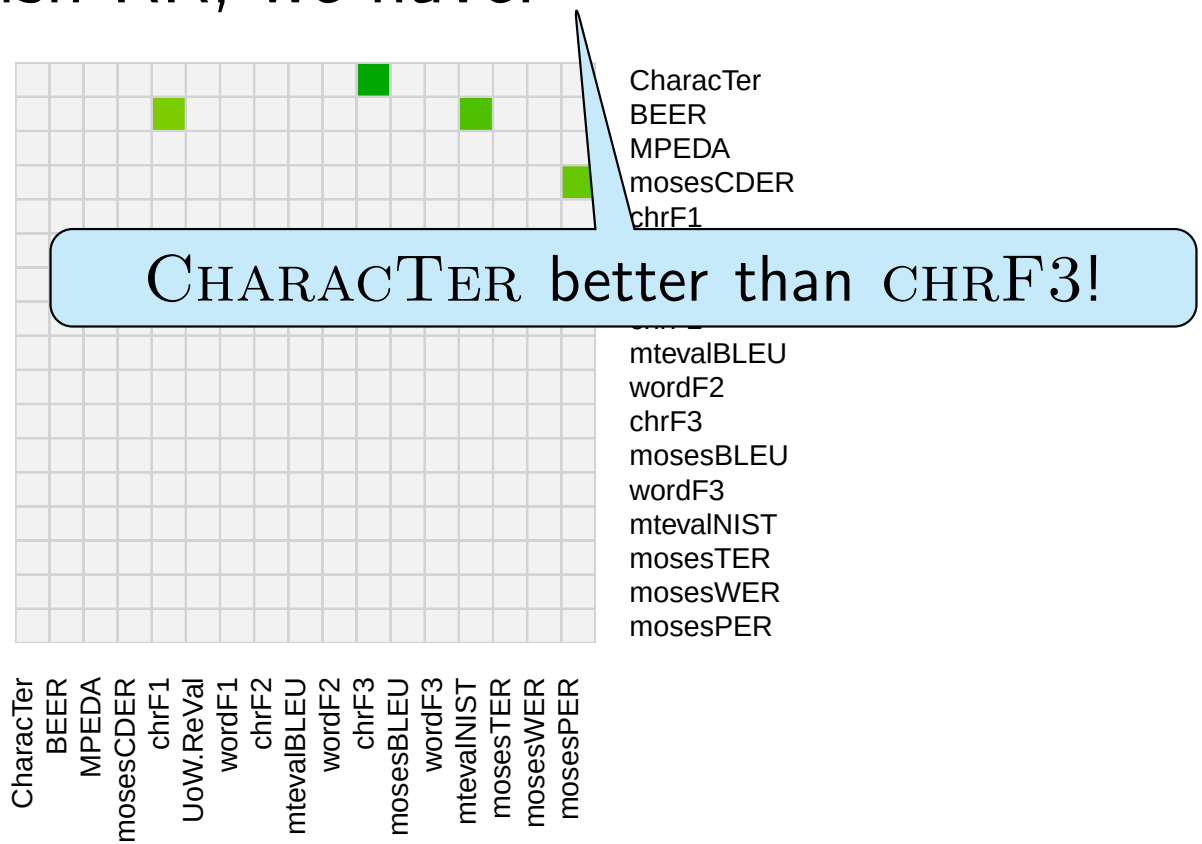
So for Czech-English RR, we have:



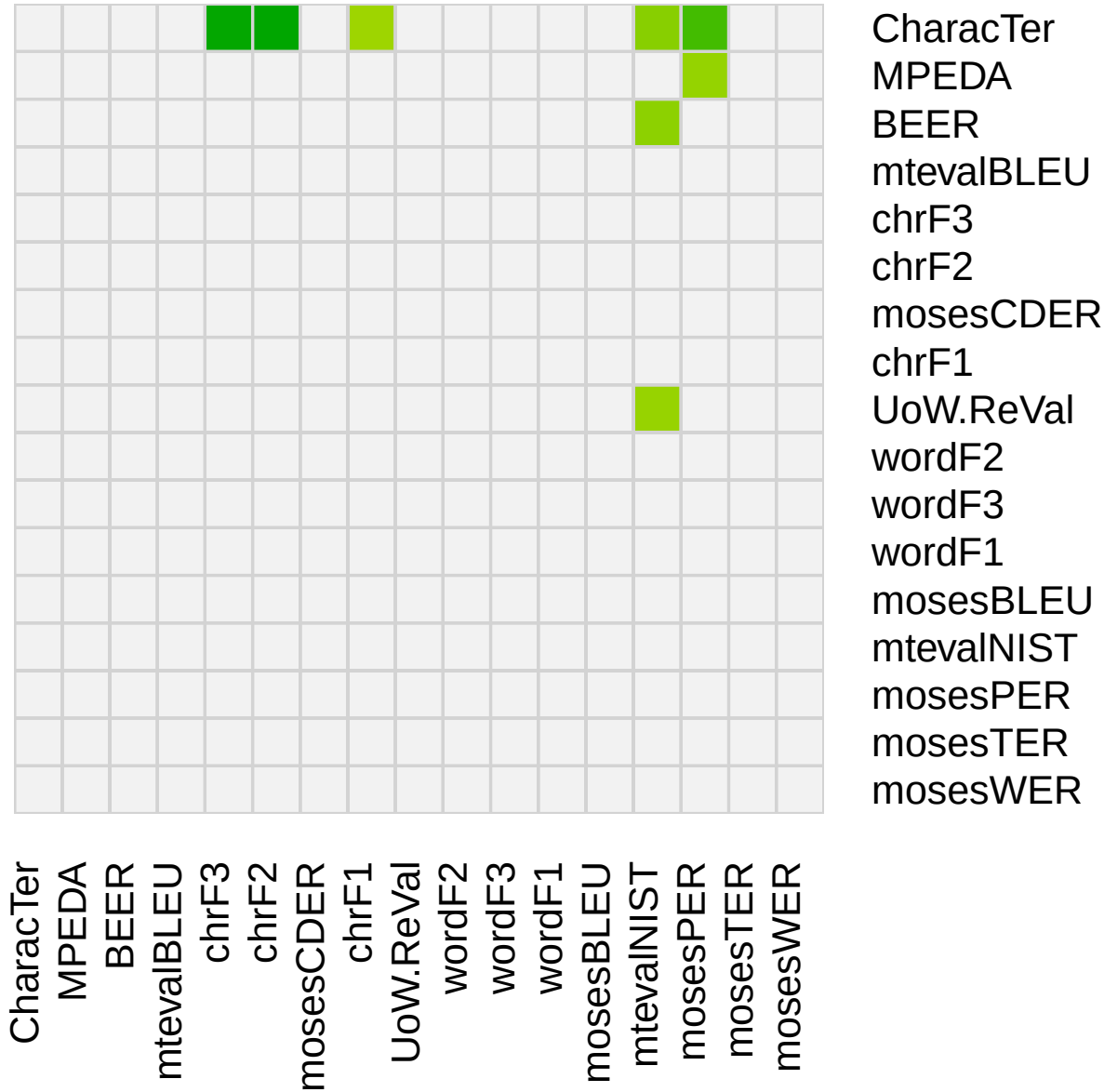
Graphical Presentation of Significant Wins

- Williams (1959) test of significant improvement in Pearson correlation.
  - Green cell indicates that the metric in the row has significantly better correlation than the metric in the column.

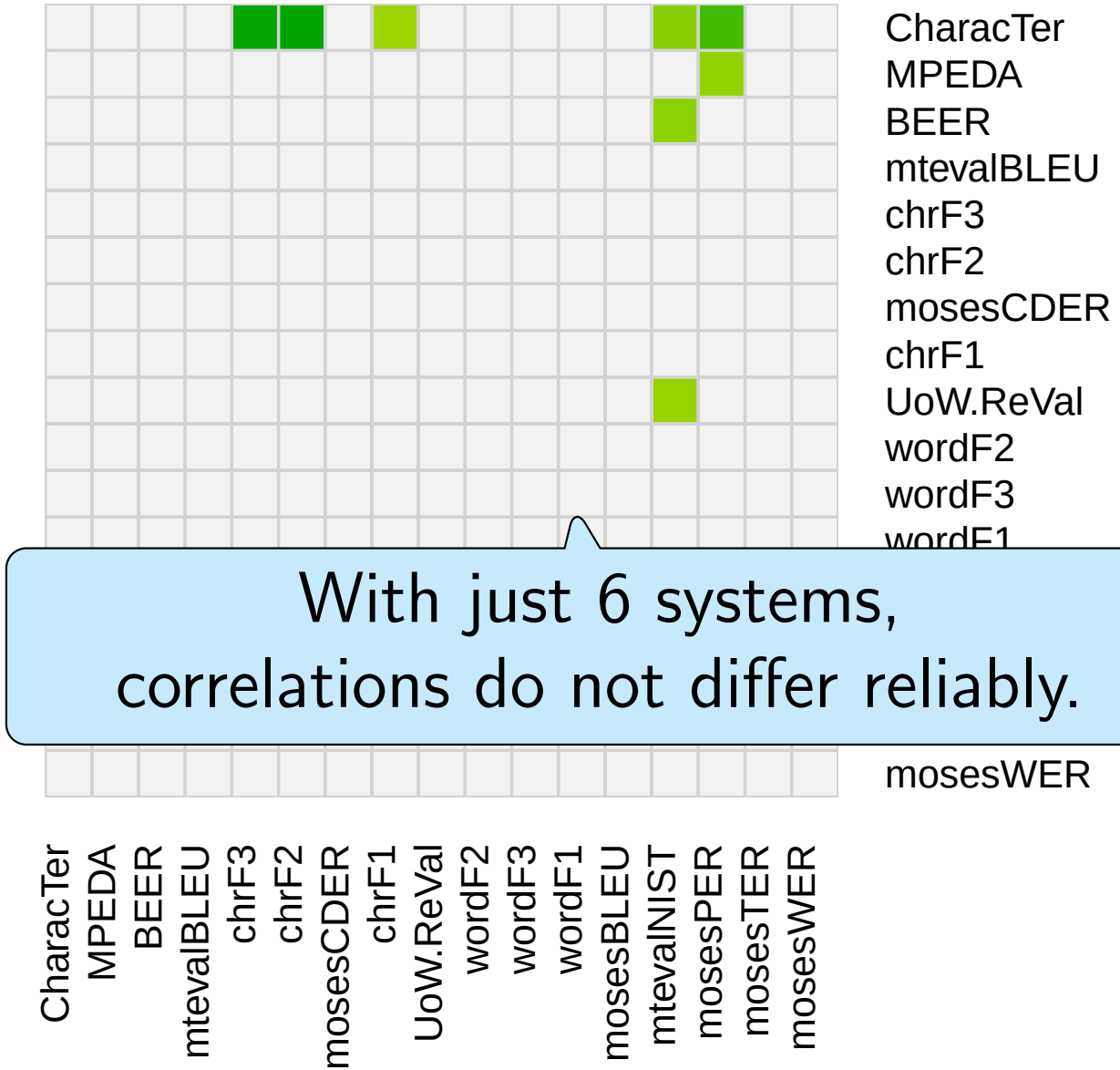
So for Czech-English RR, we have:



Czech-English Direct Assessments

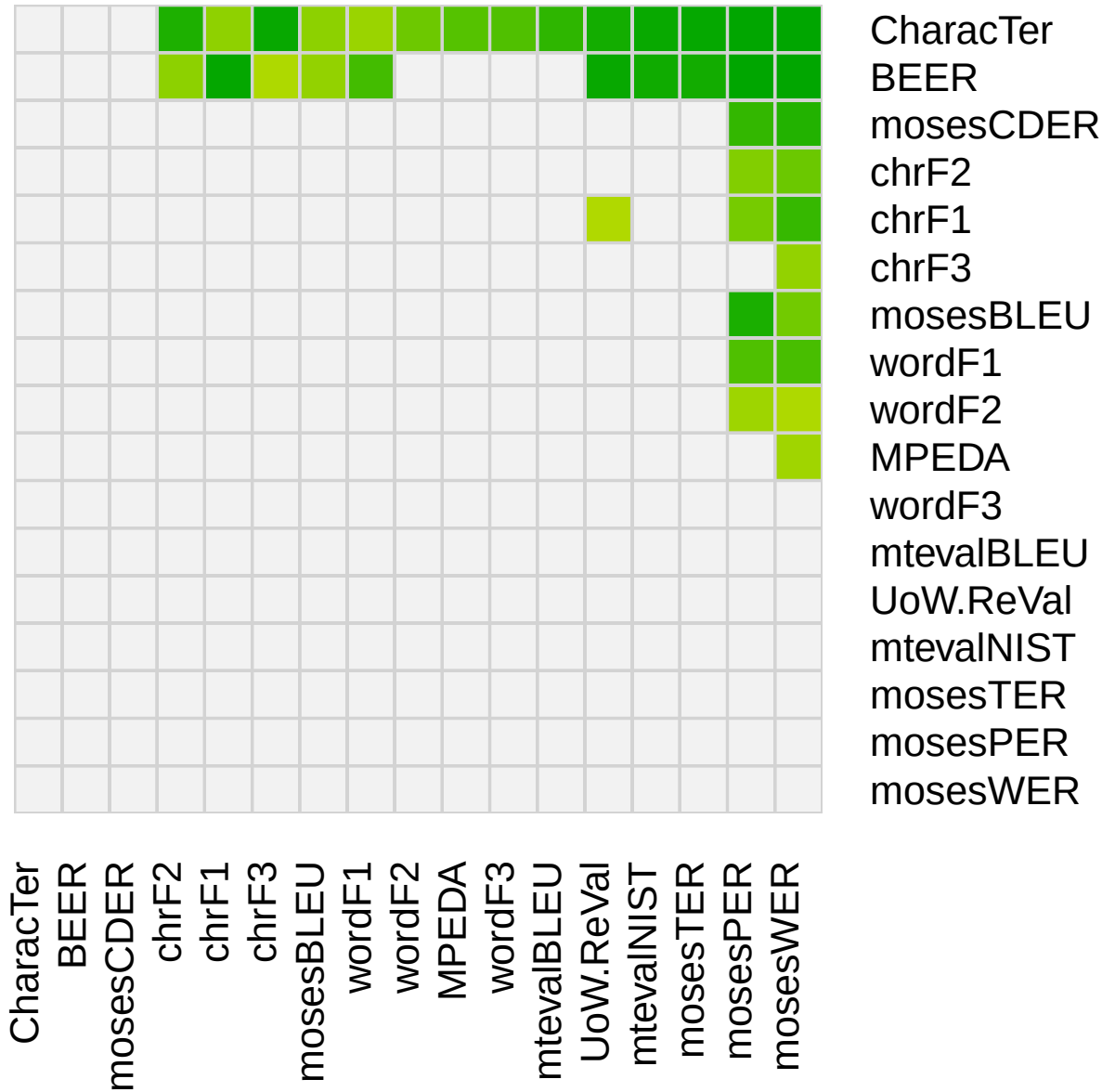


Czech-English Direct Assessments

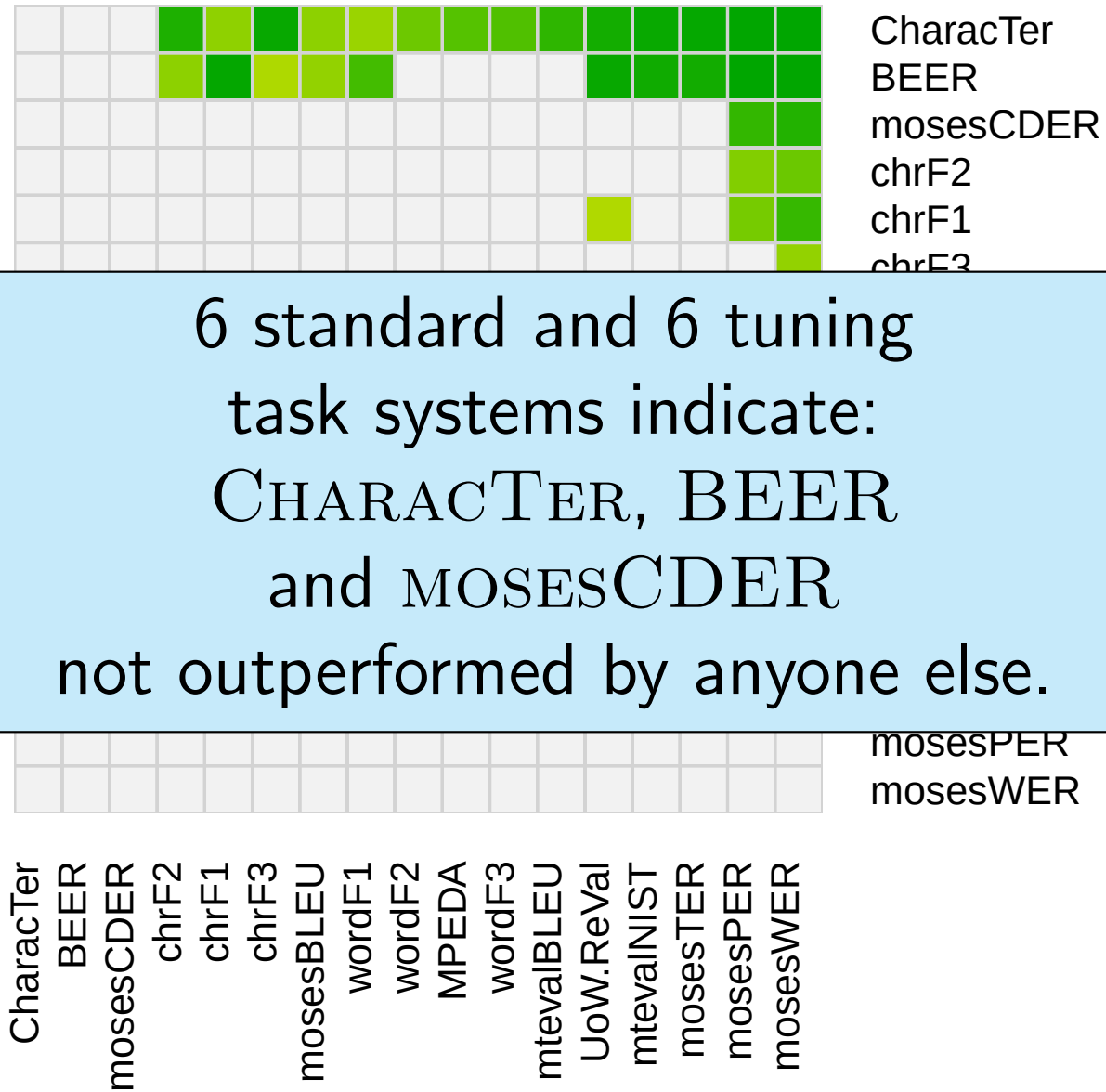


With just 6 systems,
correlations do not differ reliably.

Czech-English RR with Tuning Systems

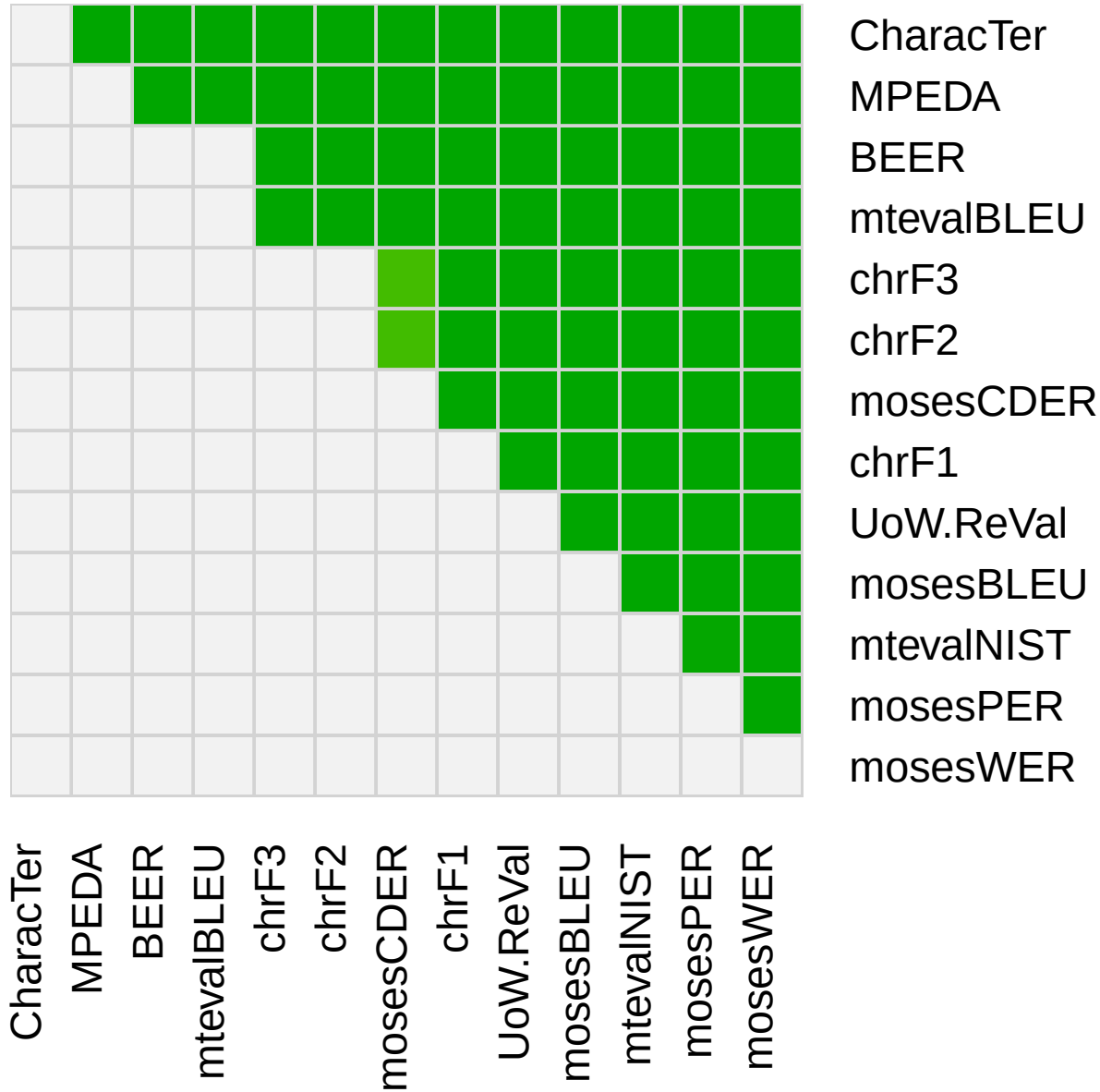


Czech-English RR with Tuning Systems

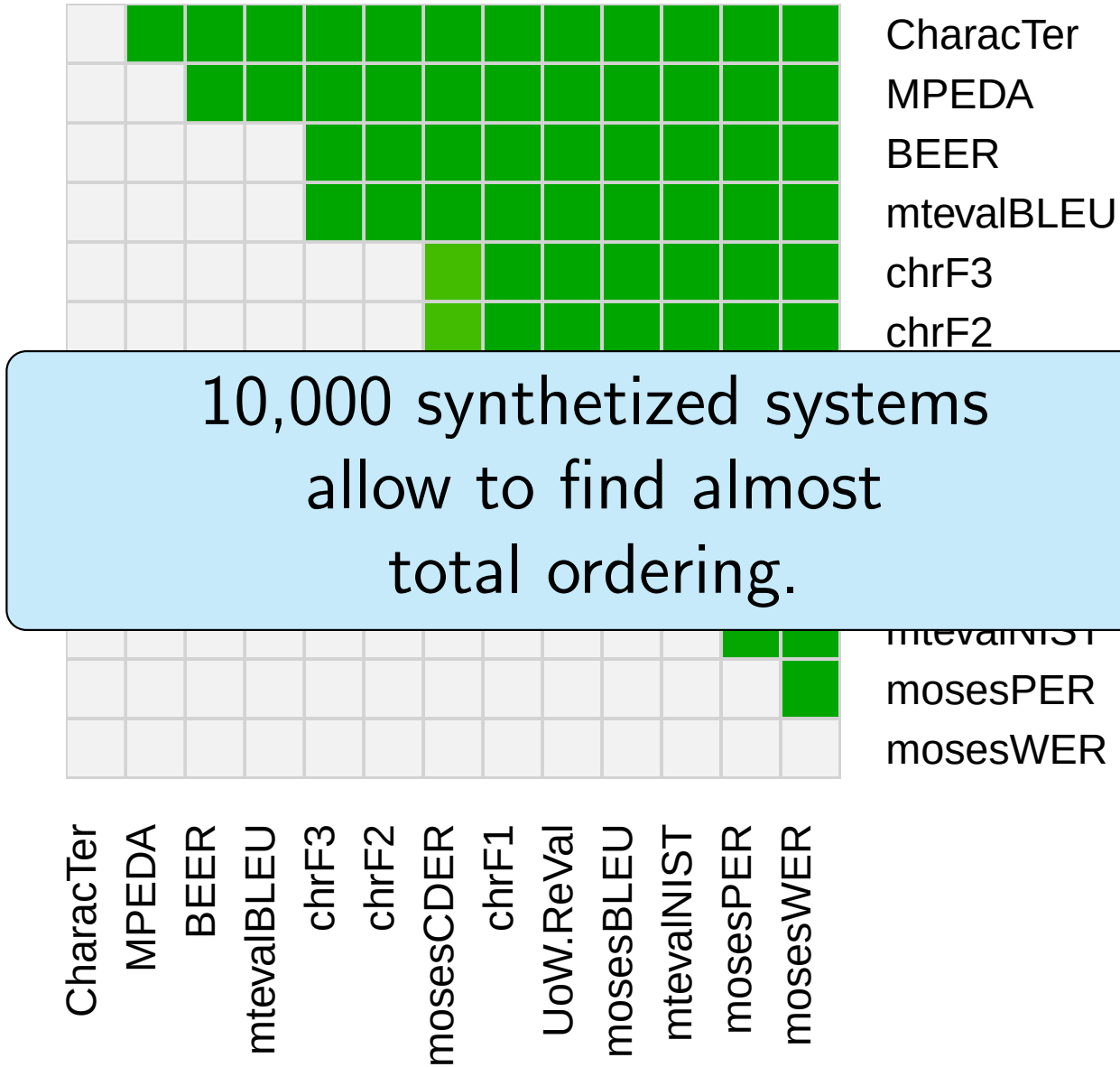


6 standard and 6 tuning task systems indicate: CharacTer, BEER and mosesCDER not outperformed by anyone else.

Czech-English DA with Hybrids



Czech-English DA with Hybrids



10,000 synthetized systems
allow to find almost
total ordering.

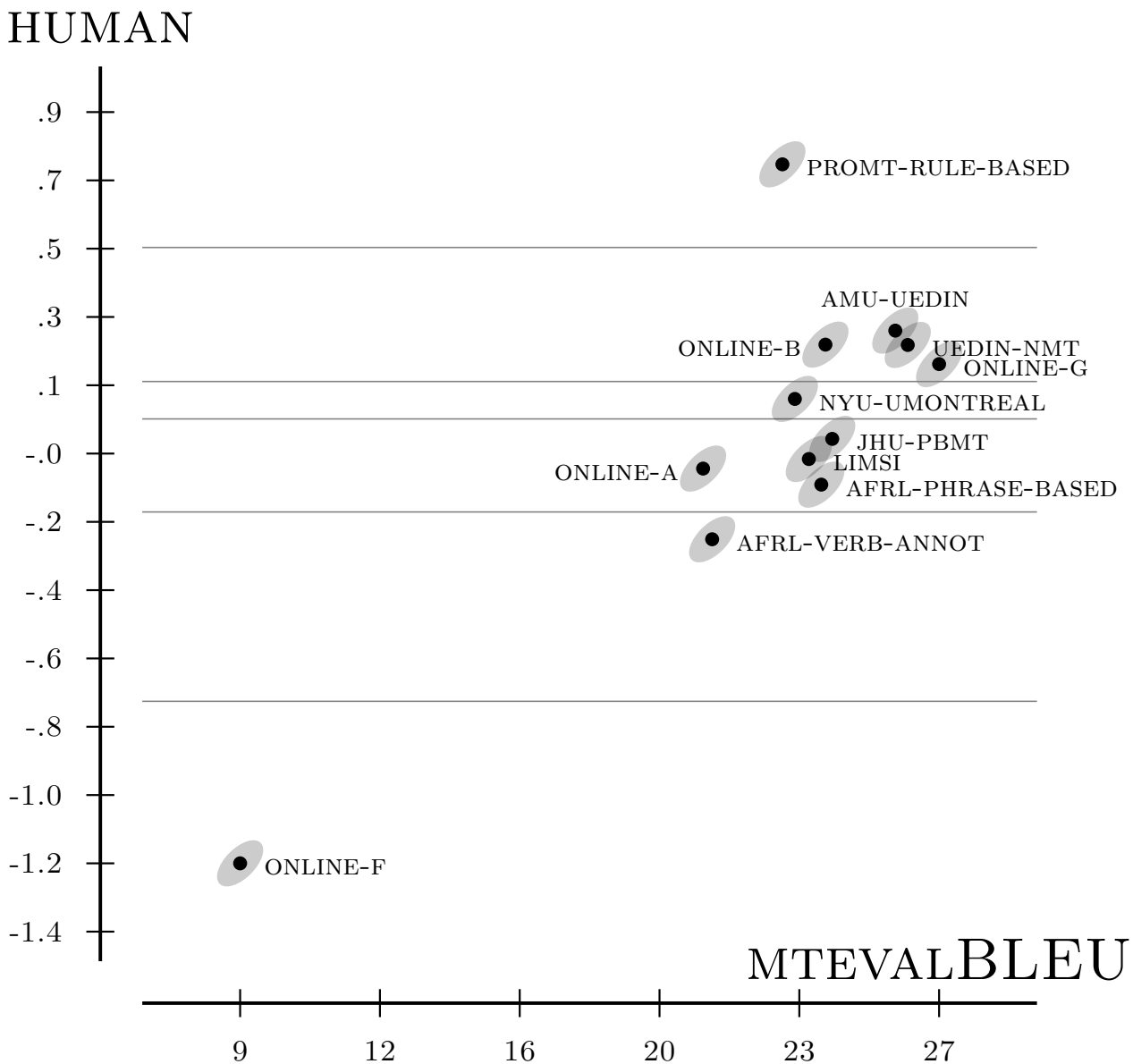
"Hybrids" = Hybrid Super-Sampling

- 10,000 "new systems" constructed by mixing sentences.
- Puts extra burden on task participants.
  - Need to score 10k "system" outputs, full test set each.
  - 200MB–1.1GB bzipped input file per language pair.
- Allows to distinguish sys-level metrics much better.
- Applicable to both RR and DA.
  - Done with DA only now because RR human judgements of individual sentences have to be carried over to these 10k systems.

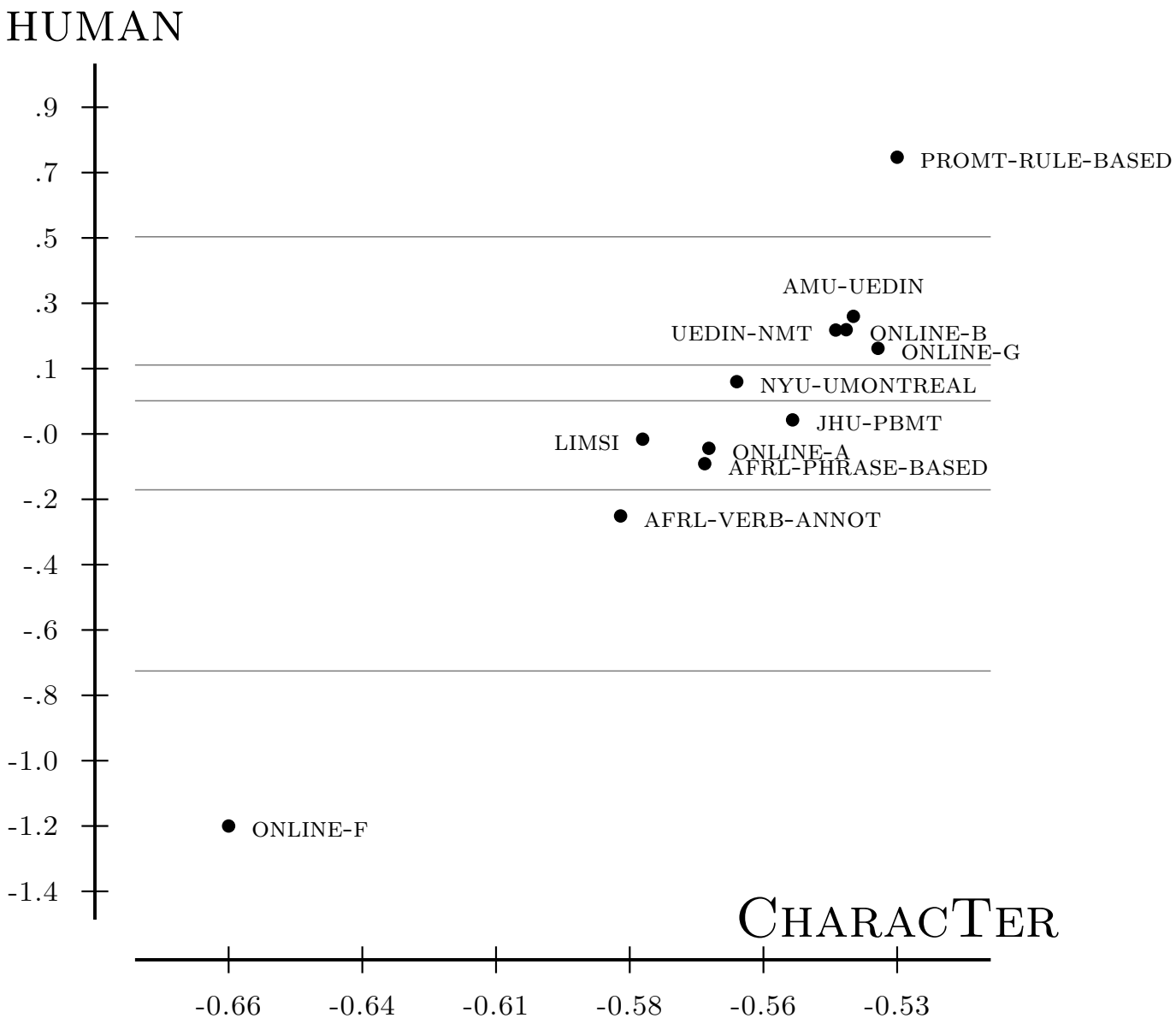
Winners according to DA hybrids:

Metric	# Wins	Language Pairs
UoW.ReVal	3	deen, roen, ruen
CharacTer	2	csen, enru
MPEDA	2	fien, tren
BEER	1	tren

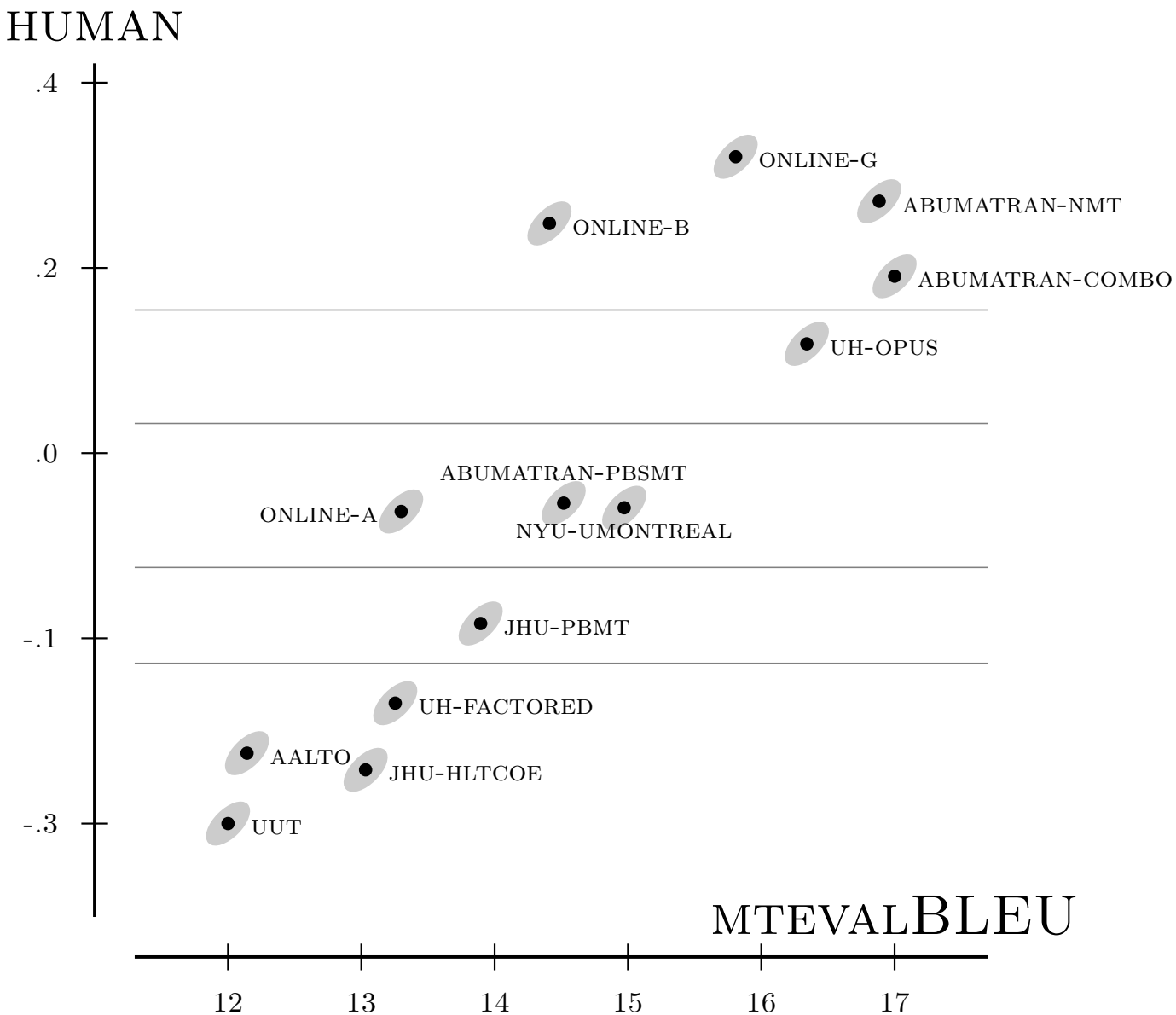
Ex. English-Russian RR BLEU



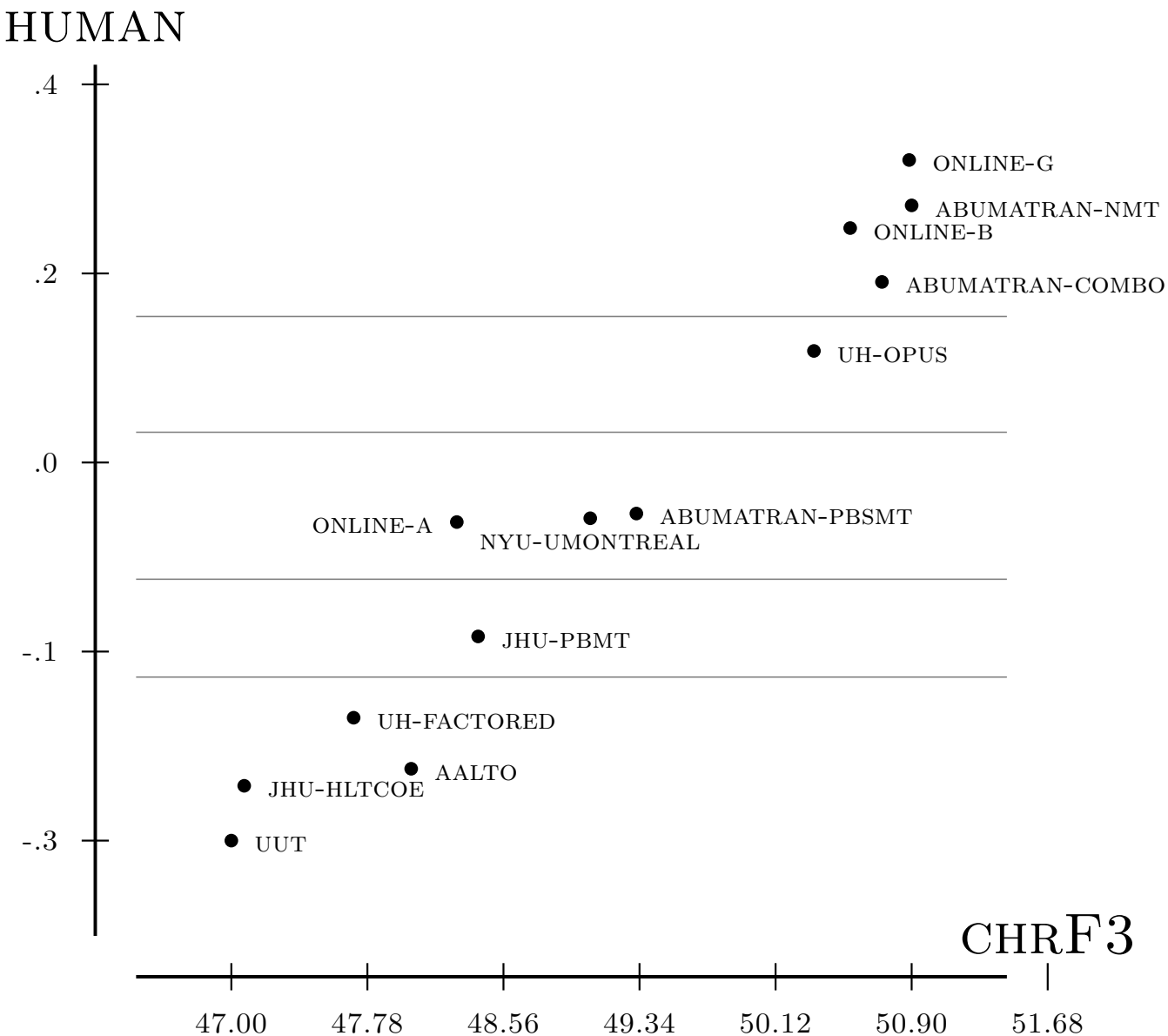
Ex. English-Russian RR CharacTer



Ex. English-Finnish RR BLEU



Ex. English-Finnish RR chrF3



Sys-Level Metrics on IT Task

- To test metrics in domain-specific setting.
- Unfortunately, often too few participating systems.

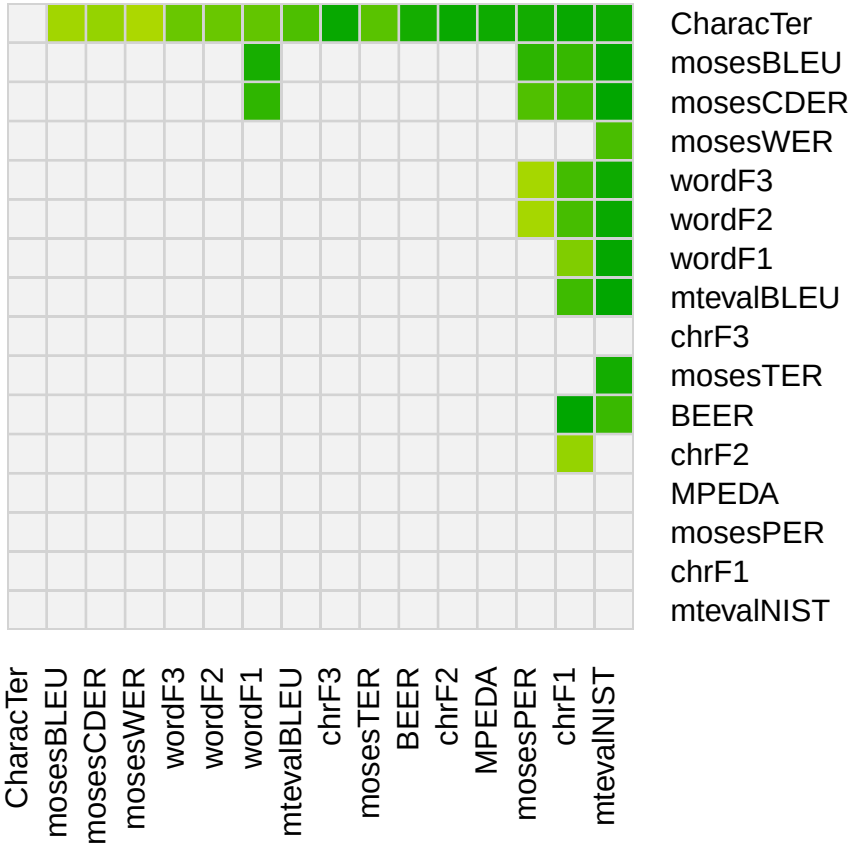
	en-bg	en-cs	en-de	en-es	en-eu	en-nl	en-pt
Human	RR	RR	RR	RR	RR	RR	RR
Systems	2	5	10	4	2	4	4
CHARACTER	**1.000**	**0.901**	**0.930**	**0.963**	**1.000**	**0.927**	0.976
CHRF3	**1.000**	**0.831**	0.700	**0.938**	**1.000**	**0.961**	0.990
CHRF2	**1.000**	0.837	0.672	0.933	**1.000**	**0.959**	0.986
BEER	**1.000**	**0.744**	0.621	0.931	**1.000**	**0.983**	**0.989**
CHRF1	**1.000**	**0.845**	0.588	0.915	**1.000**	**0.951**	0.967
MTEVALNIST	**1.000**	**0.905**	0.524	0.926	**1.000**	0.722	**0.993**
MPEDA	**1.000**	0.620	0.599	**0.951**	**1.000**	0.856	**0.989**
MOSESTER	**1.000**	**0.616**	0.628	**0.908**	**1.000**	0.835	**0.994**
MTEVALBLEU	**1.000**	**0.750**	0.621	**0.976**	**1.000**	0.596	**0.997**
MOSESWER	**1.000**	**0.009**	0.656	0.916	**1.000**	**0.903**	0.991
MOSESCDER	**1.000**	0.181	0.652	**0.932**	**1.000**	**0.914**	**0.997**
WORDF1	**1.000**	0.240	0.644	**0.959**	**1.000**	**0.911**	**0.997**
WORDF2	**1.000**	0.266	0.652	**0.965**	**1.000**	**0.900**	**0.997**
WORDF3	**1.000**	0.274	0.655	**0.966**	**1.000**	**0.897**	**0.996**
MOSESBLEU	**1.000**	0.296	0.650	**0.974**	**1.000**	0.886	**0.992**
MOSESPER	**1.000**	0.307	0.548	0.911	**1.000**	**0.938**	**0.998**

ittest2016

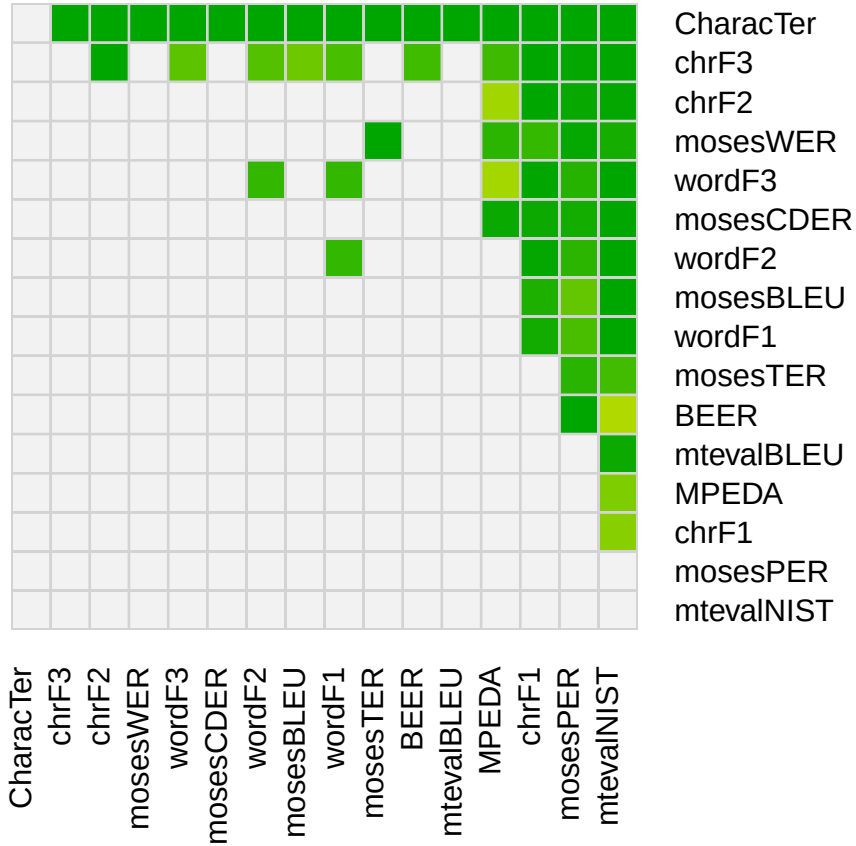
. . . so only English-German tells us something.

English-German RR News vs. IT

News
15 Systems



IT
10 Systems



- CharacTer wins in both domains.

Segment-Level News Task Evaluation

Relative Ranking	Direct Assessment
⊕ Genuine comparisons	⊖ Only 1 candidate shown
⊖ 5-way comparison hard?	⊕ Principled Pearson
⊖ Non-standard Kendall's τ	⊖ Distinct sampling needed
⊖ Conf. estimation unclear	

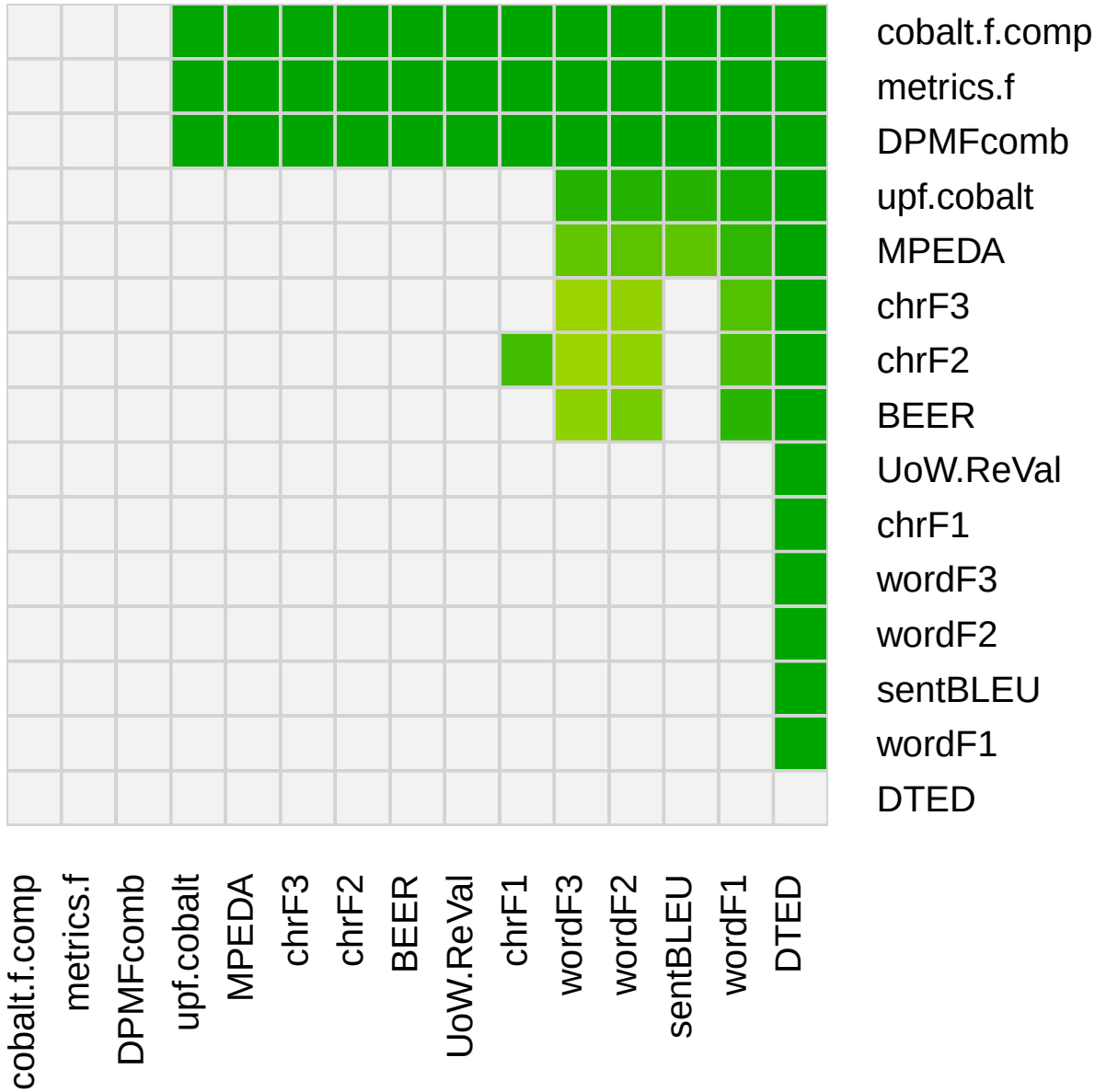
Segment-Level News Task Results

- DA and RR correlate at .85–.99 (.92 avg across langs).
- Top RR metric always among DA winners.
- RR Winners

Metric	# Wins	Language Pairs
BEER	4	encs, ende, enro, entr
DPMFcomb	3	csen, fien, ruen
metrics-f	3	deen, roen, tren
chrF2	1	enru
chrF3	1	enfi

- Williams' test for DA reveals more top-performing metrics:
  - cobalt-f (deen, ruen), MPEDA (enru)

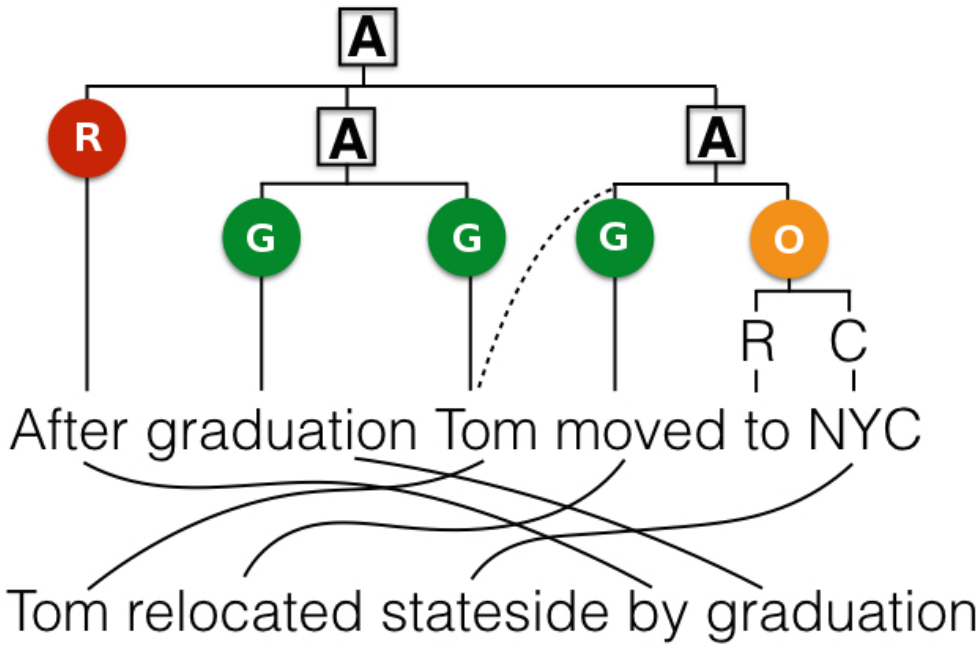
Ex. Russian-English DA Significance



Semantic Golden Truth (HUME)

HUME (Birch et al., 2016) uses two-stage annotation:

1. Semantic annotation (structure) of *source*.
2. Correctness assessment of corresponding parts of candidate.



- Final sentence-level score aggregated over source components.

HUME in Metrics Task

- A first probe.
- One test set:
  - Medical texts from Cochrane and NHS24
  - Translated by year 1 MT systems of the EU project HimL.
  - Source English annotated once.
  - Targets: Czech, German, Romanian, Polish
  - $\sim$340 sentences
- Used only in segment-level evaluation.

Results of Semantic Evaluation

Direction	**en-cs**	**en-de**	**en-ro**	**en-pl**
n	339	330	349	345
CHRF3	**.544**	**.480**	**.639**	.413
CHRF2	.537	**.479**	.634	**.417**
BEER	.516	**.480**	.620	**.435**
CHRF1	.506	**.467**	.611	**.427**
MPEDA	.468	**.478**	.595	**.425**
WORDF3	.413	.425	.587	.383
WORDF2	.408	.424	.583	.383
WORDF1	.392	.415	.569	.381
SENTBLEU	.349	.377	.550	.328

- Bold again indicates metrics not significantly outperfomed by any other (Williams, 1959).
- CHRF3 and other character-level metrics clearly win.
- SENTBLEU by far the worst.

Summary

- The **2017 golden truth** will follow the main translation task.
- Whether DA or RR, we will **use hybrids** for sys-level.
- Domain-specific evaluation of metrics needs **enough systems** to participate (or plan seg-level evaluation).
- Top metrics consider again **character sequences** and are **trained**.
- Even "semantics" seems well captured by **character-level metrics**.

References

Alexandra Birch, Barry Haddow, Ondřej Bojar, and Omri Abend. 2016. Hume: Human ucca-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030* .

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. 2016. CobaltF: A Fluent Metric for MT Evaluation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal.

Martin McCaffery and Mark-Jan Nederhof. 2016. DTED: Evaluation of Machine Translation Structure Using Dependency Parsing and Tree Edit Distance. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisboa, Portugal.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany.

Evan James Williams. 1959. *Regression analysis*, volume 14. Wiley New York.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisboa, Portugal.