Findings of the 2016 Conference on Machine Translation

WMT 2016 @ ACL
Berlin, Germany
August 11–12

***Organizers**: Ondřej Bojar (Charles University in Prague), Christian Buck (University of Edinburgh), Rajen Chatterjee (FBK), Christian Federmann (MSR), Liane Guillou (University of Edinburgh), Barry Haddow (University of Edinburgh), Matthias Huck (University of Edinburgh), Antonio Jimeno Yepes (IBM Research Australia), Varvara Logacheva (University of Sheffield), Aurélie Névéol (LIMSI, CNRS), Mariana Neves (Hasso-Plattner Institute), Pavel Pecina (Charles University in Prague), Martin Popel (Charles University in Prague), Philipp Koehn (University of Edinburgh / Johns Hopkins University), Christof Monz (University of Amsterdam), Matteo Negri (FBK), Matt Post (Johns Hopkins University), Carolina Scarton (University of Sheffield), Lucia Specia (University of Sheffield), Karin Verspoor (University of Melbourne), Jörg Tiedemann (University of Helsinki), Marco Turchi (FBK)*

News Translation Task

Overview

~~Français~~

čeština
Deutsch
NEW română
рýсский
suomi
NEW Türkçe

English

Funding

- European Union's Horizon 2020 program

- Yandex (Russian–English and Turkish–English test sets)
- University of Helsinki (Finnish–English test set)

Participation

102 entries from 24 institutions

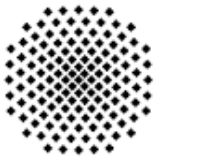
+4 anonymized commercial, online, and rule-based systems

















Universität Stuttgart

Aalto University

Carnegie Mellon University

Université de Montréal

PROMT TRANSLATOR

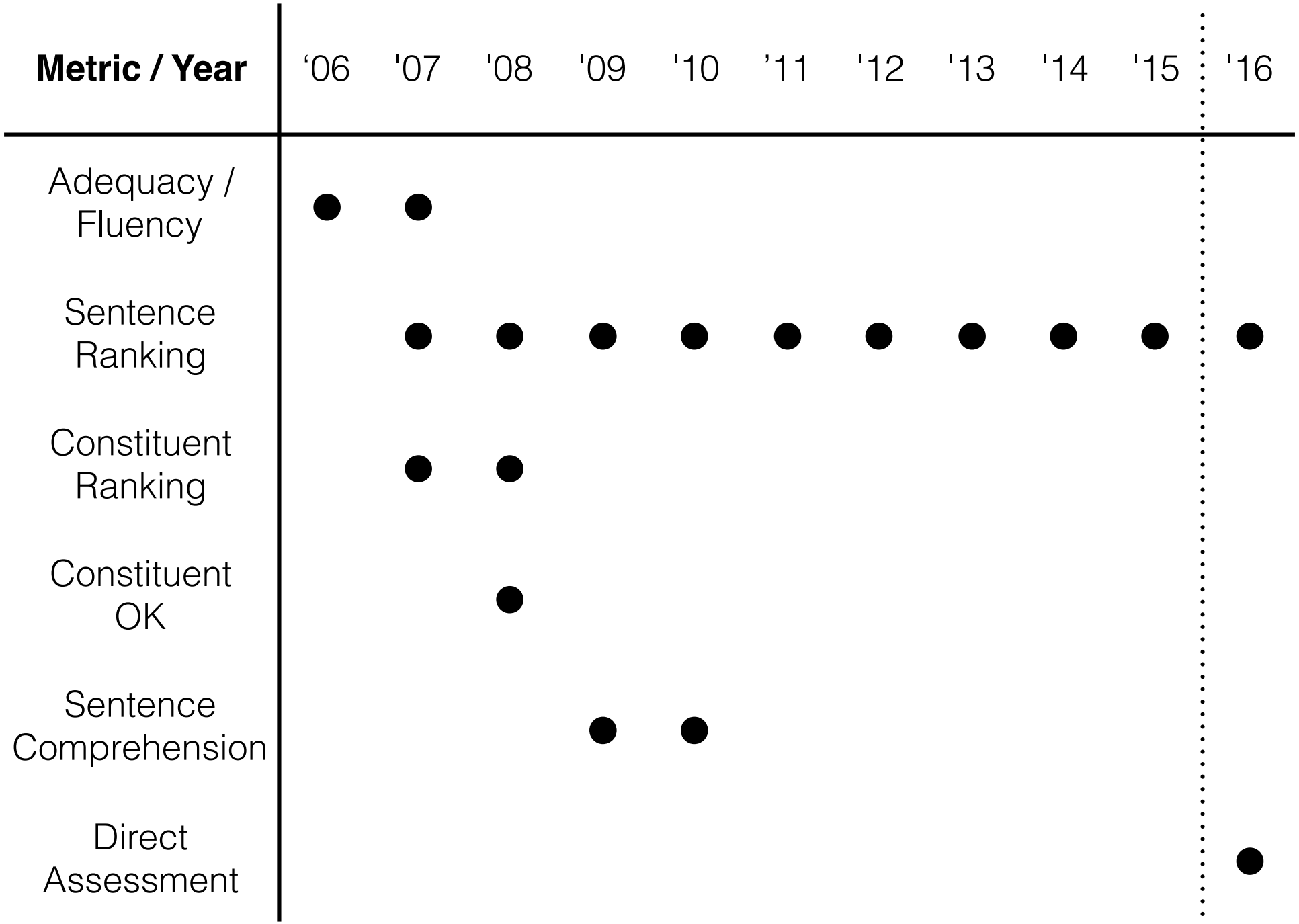
RWTH AACHEN UNIVERSITY

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

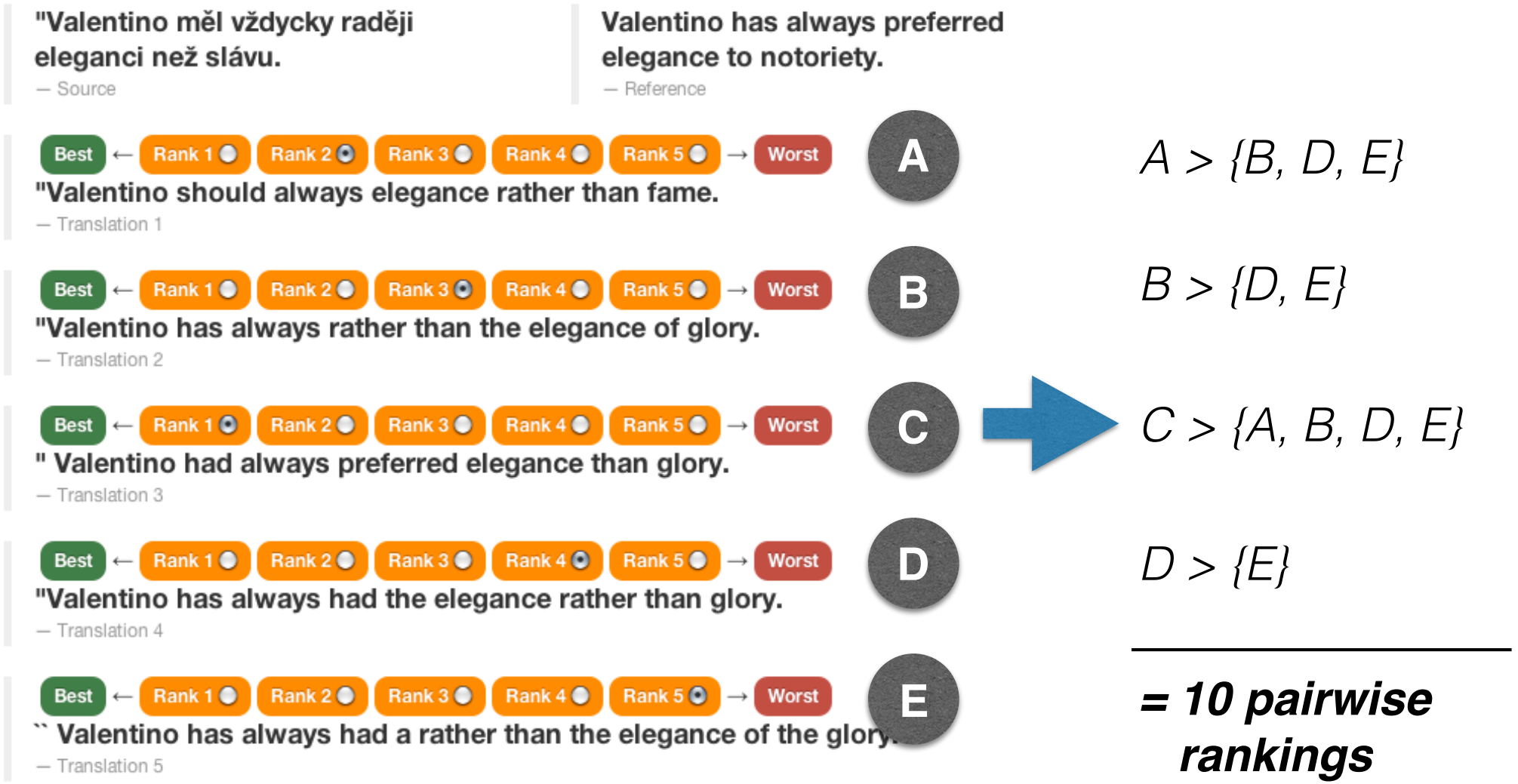
Human Evaluation

Human Evaluation

- We wish to identify the best systems for each task
  - Automatic metrics are useful for development, but must be grounded in **human evaluation** of system output
- How to compute it?
  - Adequacy / fluency, **sentence ranking (RR)**, constituent ranking, constituent OK, sentence comprehension
  - **Direct Assessment (DA)**


Metric / Year
‘06
'07
'08
'09
'10
’11
'12
'13
'14
'15
'16
Adequacy / Fluency
Sentence Ranking
Constituent Ranking
Constituent OK
Sentence Comprehension
Direct Assessment

Sentence Ranking

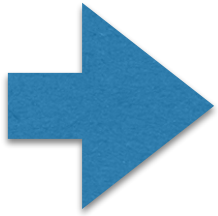


https://github.com/cfedermann/Appraise/

More Judgments

- Innovation: rank distinct outputs instead of systems

	1	2	3	4	5
F				•	
A				•	
B		•			
J					•
H			•		



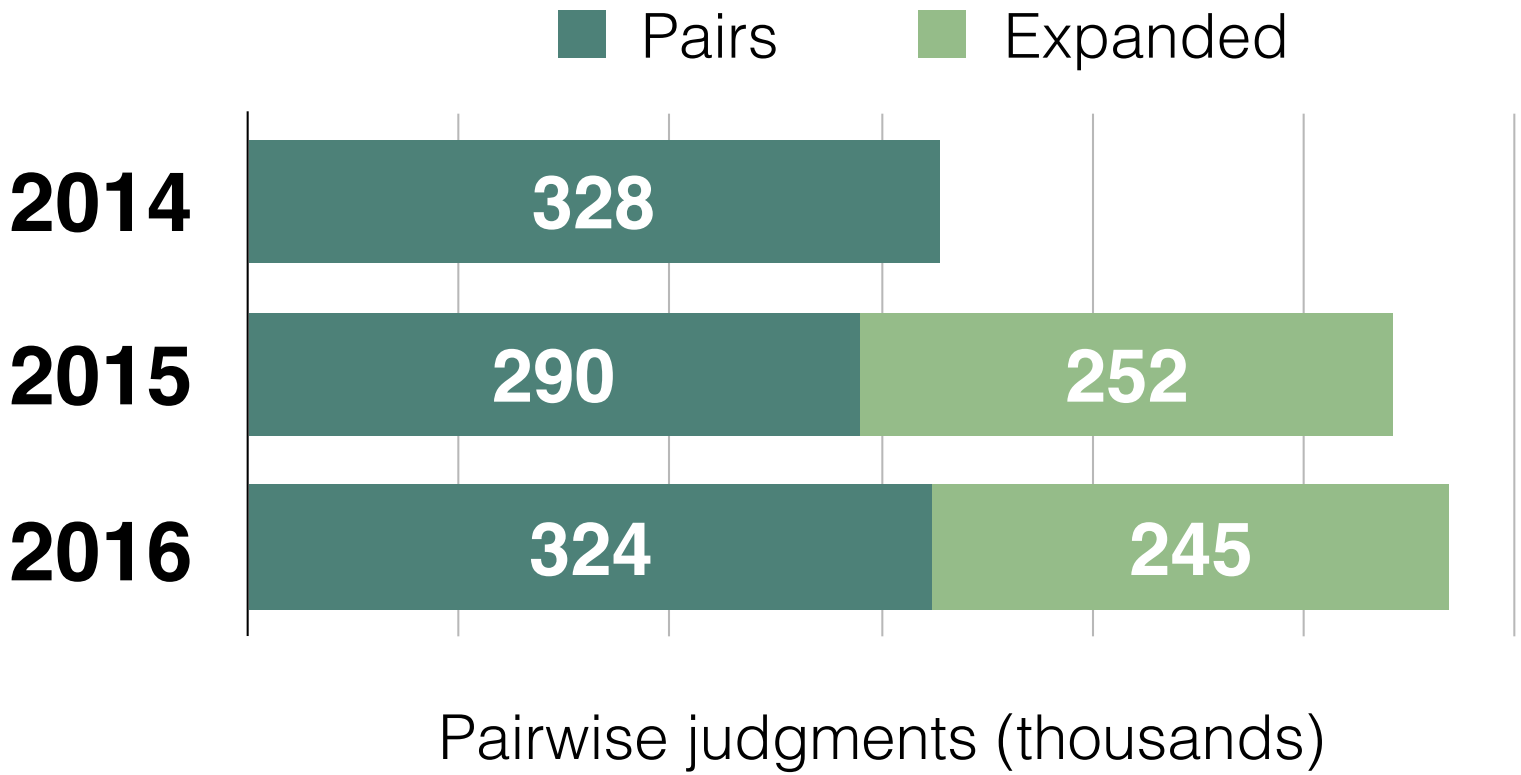
	1	2	3	4	5
AF				•	
B		•			
J					•
H			•		

- Then, distribute rankings across systems:

$$A > B, A = F, A > H, A < J$$
$$B < F, B < H, B < J$$
$$F > H, F < J$$
$$H < J$$

Data collected

- 150 trusted annotators, 939 person-hours

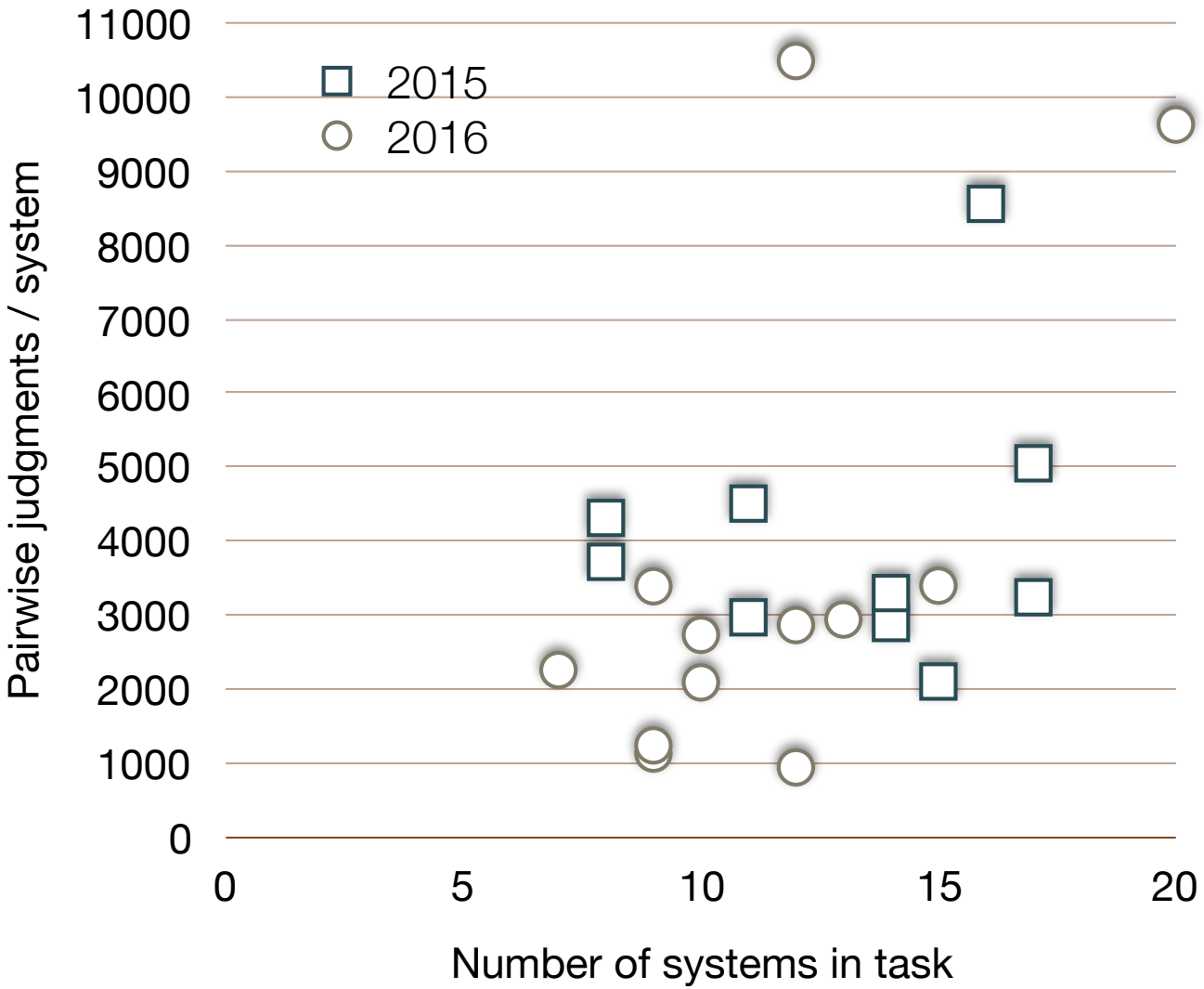


statmt.org/wmt16/results.html

Clustering

- Rank systems using TrueSkill (Herbrich et al., 2006, Sakaguchi et al., 2014)
- Cluster (Koehn, 2012)
  - Aggregate each system's rank over 1,000 bootstrap-resampled folds
  - Throw out top and bottom 25 ranks, collect ranges
  - Groups systems by non-overlapping ranges

Manual evaluation summary



- ~4.1k rankings / task (~3k last year)
- Total judgments: 542k (328k last year)
- Data: statmt.org/wmt16/results.html

Czech–English

cluster	constrained	not constrained
1	uedin-nmt	
2	jhu-pbmt	
3		online-B
4	PJATK, TT-*	
5		online-A
6	cu-mergetrees	

English–Czech

cluster	constrained	not constrained
1	uedin-nmt	
2	nyu-montreal	
3	jhu-pbmt	
4	cu-chimera, cu-tamchyna	
5	uedin-cu-syntax	online-B
6	TT-*	
7		online-A
8	cu-tectomt	
9	tt-usaar-hmm-mert	
10	cu-mergetrees	
11	tt-usaar-hmm-mira	
12	tt-usaar-harm	

Russian–English

cluster	constrained	not constrained
1	amu-uedin,NRC, uedin-nmt	online-G, online-B
2	AFRL-MITLL-phr	online-A
3	AFRL-MITLL-cntr, PROMT-rule	
4		online-F

English–Russian

cluster	constrained	not constrained
1		**promt-rule**
2	amu-uedin, uedin-nmt	online-B, online-G
3	NYU-montreal	
4	jhu-pbmt, limsi, AFRL-MITLL-phr	online-A
5	AFRL-MITLL-verb	
6		online-F

German–English

cluster	constrained	not constrained
1	uedin-nmt	
2	uedin-syntax, kit, uedin-pbmt, jhu-pbmt	online-B, online-A
3	jhu-syntax	online-G
4		online-F

English–German

cluster	constrained	not constrained
1	uedin-nmt	
2	metamind	
3	uedin-syntax	
4	nyu-montreal	
5	kit-limsi, cambridge, promt-rule, kit	online-B, online-A
6	jhu-syntax, jhu-pbmt	
7	uedin-pbmt	online-F, online-G

Romanian–English

cluster	constrained	not constrained
1	uedin-nmt	online-B
2	uedin-pbmt	
3	uedin-syntax, jhu-pbmt, limsi	online-A

English–Romanian

cluster	constrained	not constrained
1	uedin-nmt, qt21-himl-comb	
2	kit, uedin-pbmt, uedin-lmu-hiero, rwth-comb	online-B
3	limsi, lmu-cuni, jhu-pbmt, usfd-rescoring	online-A

Finnish–English

cluster	constrained	not constrained
1		uedin-pbmt, online-G, online-B, uh-opus
2		PROMT-smt
3	uh-factored, uedin-syntax	
4		online-A
5	jhu-pbmt	

English–Finnish

cluster	constrained	not constrained
1	abumatran-nmt, abumatran-cmb	online-G, online-B, uh-opus
2	abumatran-pb, nyu-montreal	online-A
3	jhu-pbmt, uh-factored, aalto, jhu-hltcoe, uut	

Turkish–English

cluster	constrained	not constrained
1		online-B, online-G, online-A
2	tbtk-syscomb, usda	PROMT-smt
3	jhu-syntax, jhu-pbmt, parFDA	

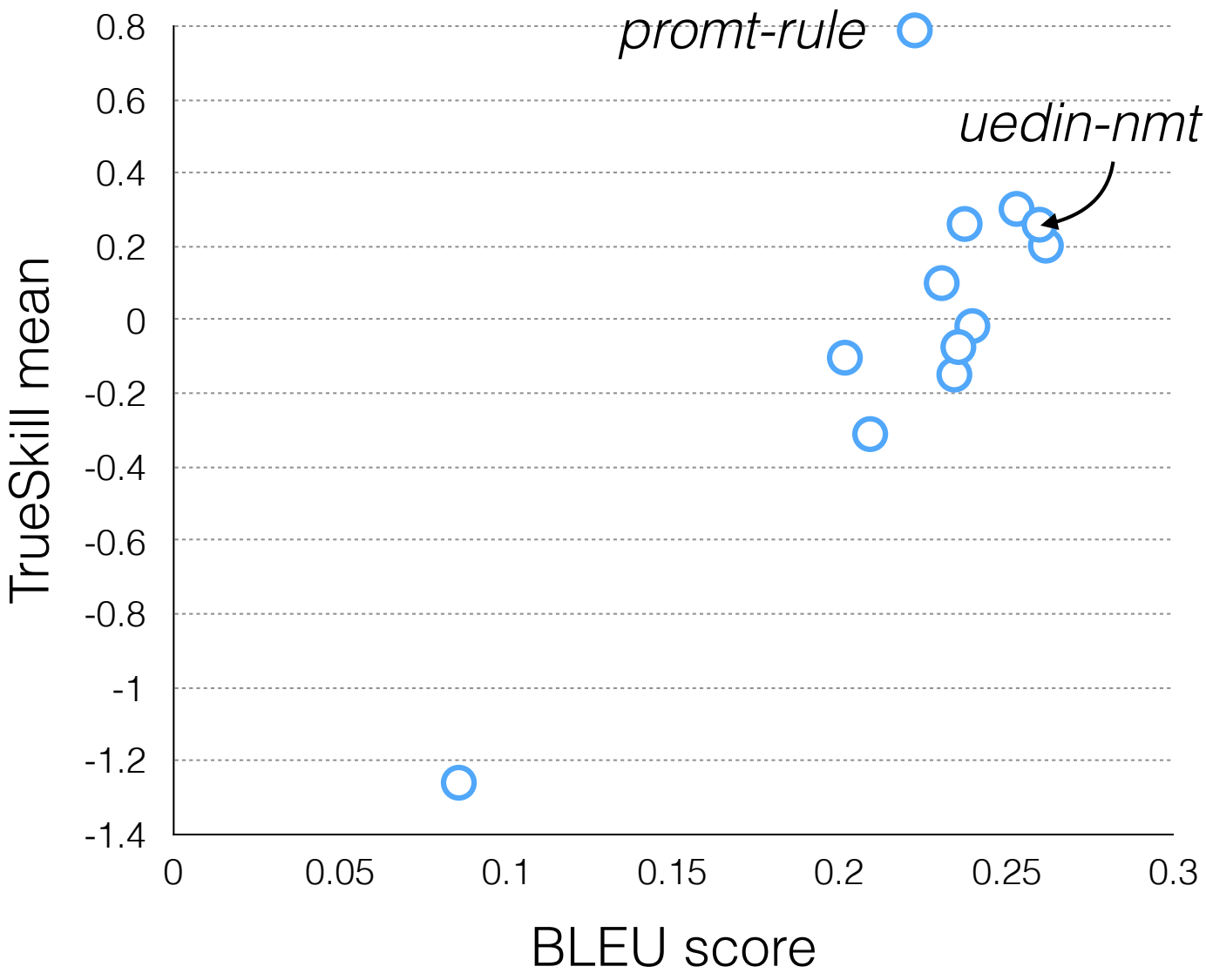
English–Turkish

cluster	constrained	not constrained
1		online-G, online-B
2		online-A
3	ysda	
4	jhu-hltcoe, tbtk-morph, cmu	
5	jhu-pbmt, parFDA	

Trends

- UEdin-NMT
  - 4 languages: uncontested winner
  - 3 languages: tied for first
  - 1 language: tied for second (behind rule-based!)
- English–Russian: rule-based system (PROMT-rule) the winner by a wide margin

Comparison with BLEU



Data

- statmt.org/wmt16/results.html
  - Source and reference data, system outputs
  - Manual evaluation results (raw XML, CSV files with pairwise rankings)

```
srclang,trglang,id,judge,sys1,sys1rank,sys2,sys2rank,group
deu,eng,348,judge13,jhu-syntax,3,online-B,5,190
```

- github.com/cfedermann/wmt16
  - Code used to compute rankings, clusters, annotator agreement

Direct Assessment