Findings of the 2015 Workshop on Statistical Machine Translation

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Mateo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi

WMT 2015 @ EMNLP
Lisbon, Portugal
September 17–18

Human Evaluation

- We wish to identify the best systems for each task

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

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?

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**, constituent ranking, constituent OK, sentence comprehension

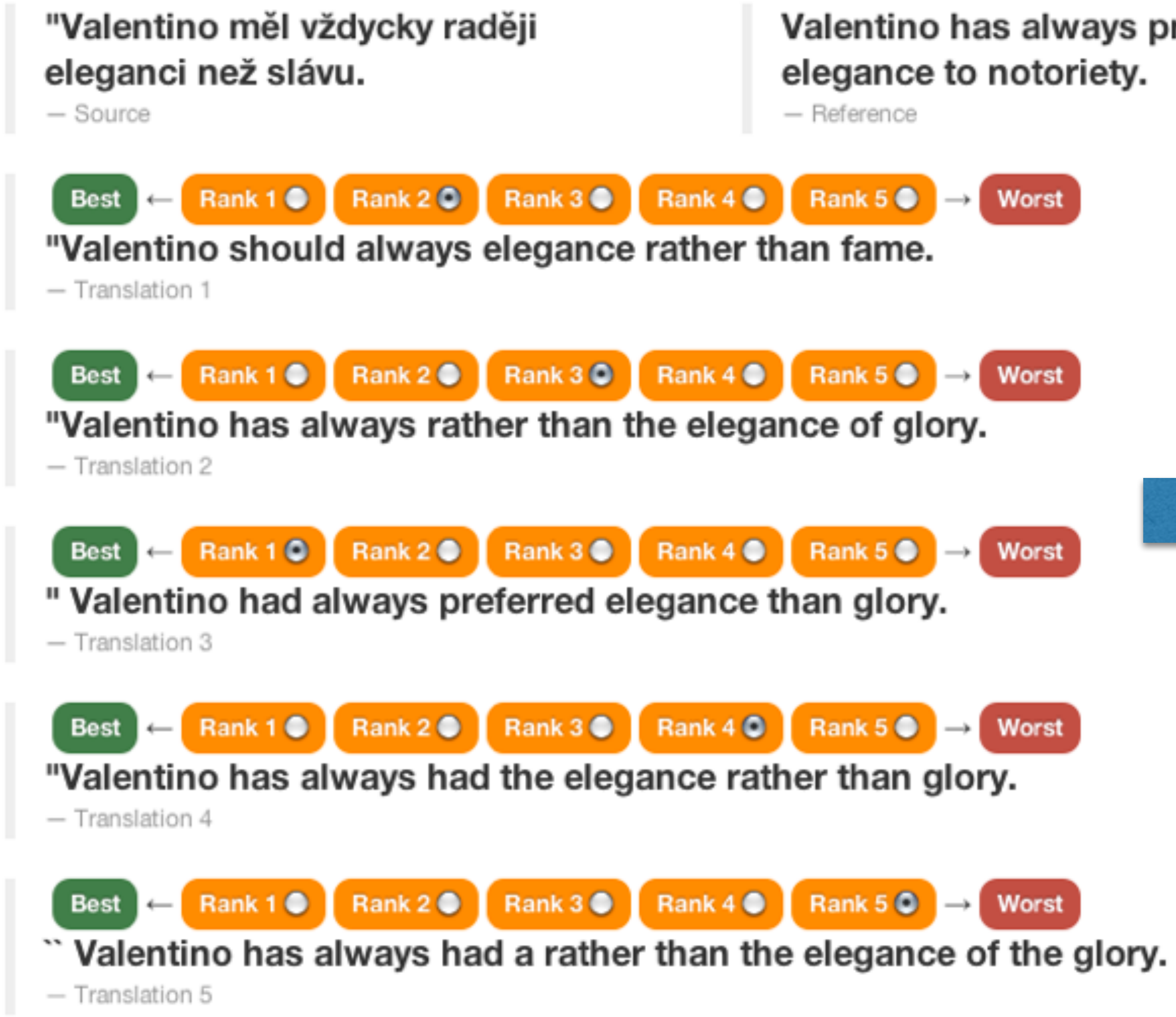
Metric / Year	‘06	'07	'08	'09	'10	’11	'12	'13	'14	'15
Adequacy / fluency	●	●								🟦
Sentence ranking		●	●	●	●	●	●	●	●	🟦
Constituent ranking		●	●							🟦
Const OK (Y/N)			●							🟦
Sentence comprehension				●	●					🟦

slide due to Ondrej Bojar

Metric / Year	‘06	'07	'08	'09	'10	’11	'12	'13	'14	'15
Adequacy / fluency	●	●								
Sentence ranking		●	●	●	●	●	●	●	●	●
Constituent ranking		●	●							
Const OK (Y/N)			●							
Sentence comprehension				●	●					

slide due to Ondrej Bojar

Sentence Ranking



"Valentino měl vždycky raději eleganci než slávu.
— Source

Valentino has always preferred elegance to notoriety.
— Reference

Best ← Rank 1 ○ Rank 2 ● Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
"Valentino should always elegance rather than fame.
— Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ● Rank 4 ○ Rank 5 ○ → Worst
"Valentino has always rather than the elegance of glory.
— Translation 2

Best ← Rank 1 ● Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
" Valentino had always preferred elegance than glory.
— Translation 3

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ● Rank 5 ○ → Worst
"Valentino has always had the elegance rather than glory.
— Translation 4

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ● → Worst
`` Valentino has always had a rather than the elegance of the glory.
— Translation 5

A > {B, D, E}

B > {D, E}

C > {A, B, D, E}

D > {E}

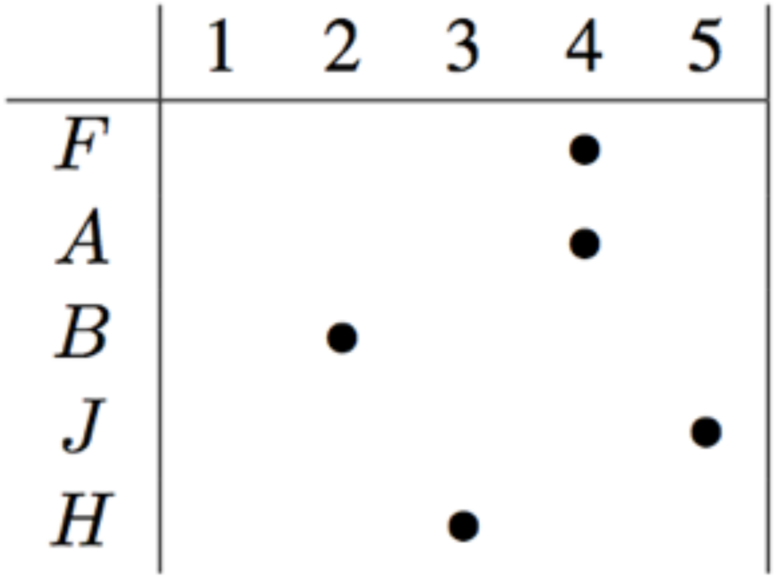
= 10 pairwise rankings

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

More Judgments

More Judgments

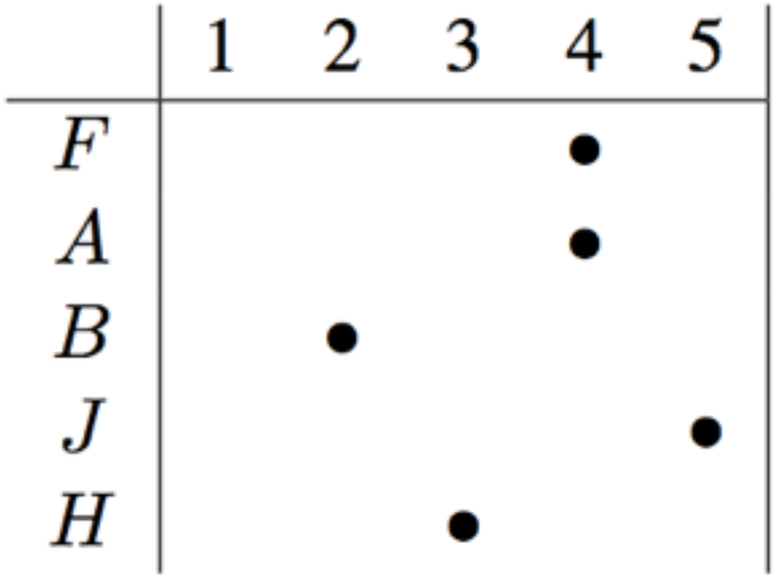
- Innovation: rank distinct outputs instead of systems



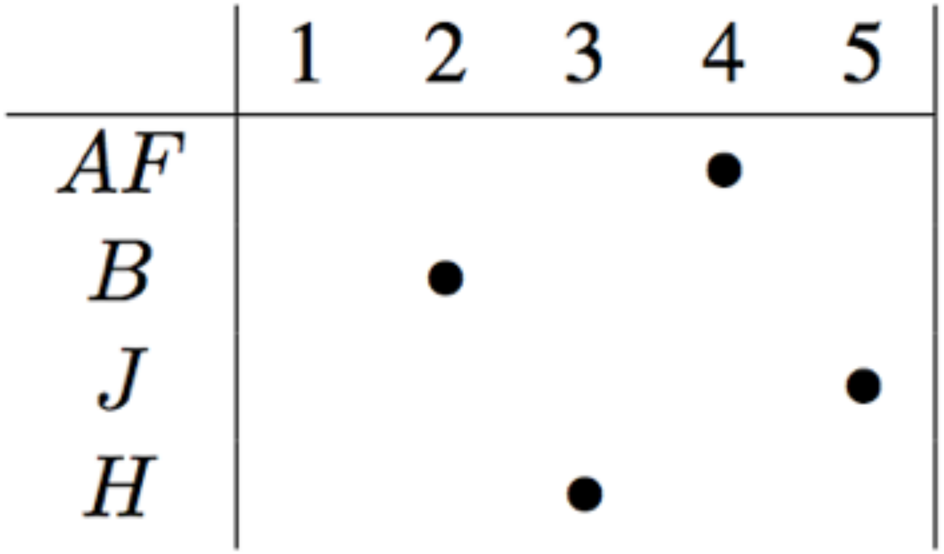
	1	2	3	4	5
F				●	
A				●	
B		●			
J					●
H			●		

More Judgments

- Innovation: rank distinct outputs instead of systems



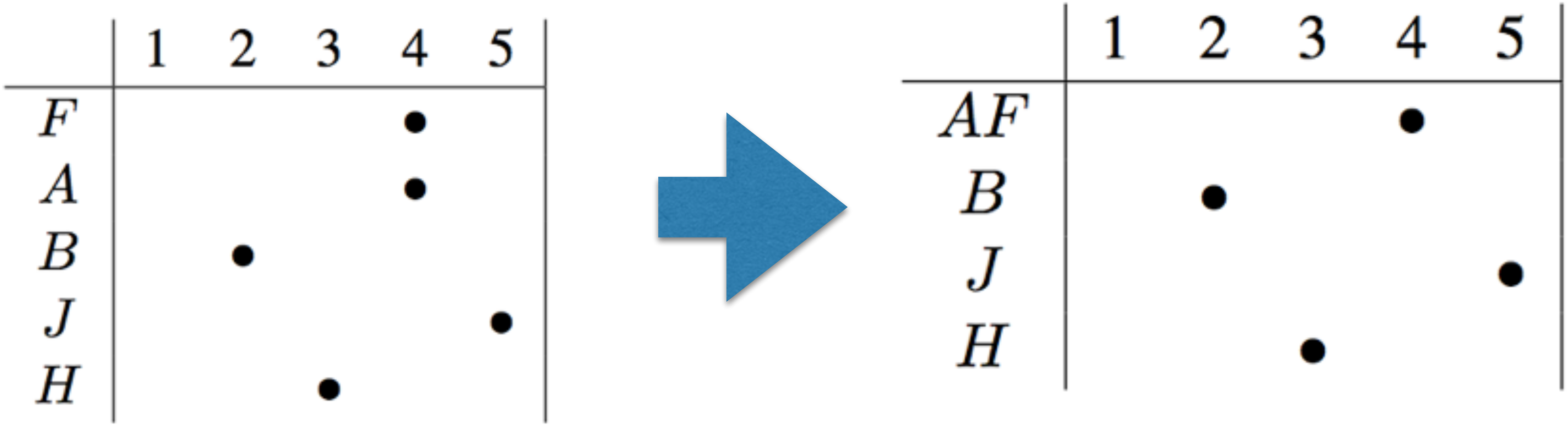
	1	2	3	4	5
F				●	
A				●	
B		●			
J					●
H			●		



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

More Judgments

- Innovation: rank distinct outputs instead of systems



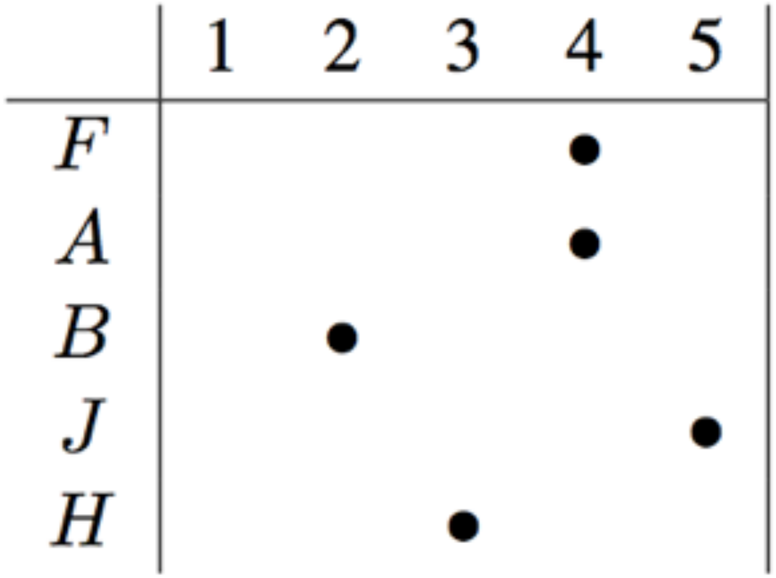
	1	2	3	4	5
F				●	
A				●	
B		●			
J					●
H			●		

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

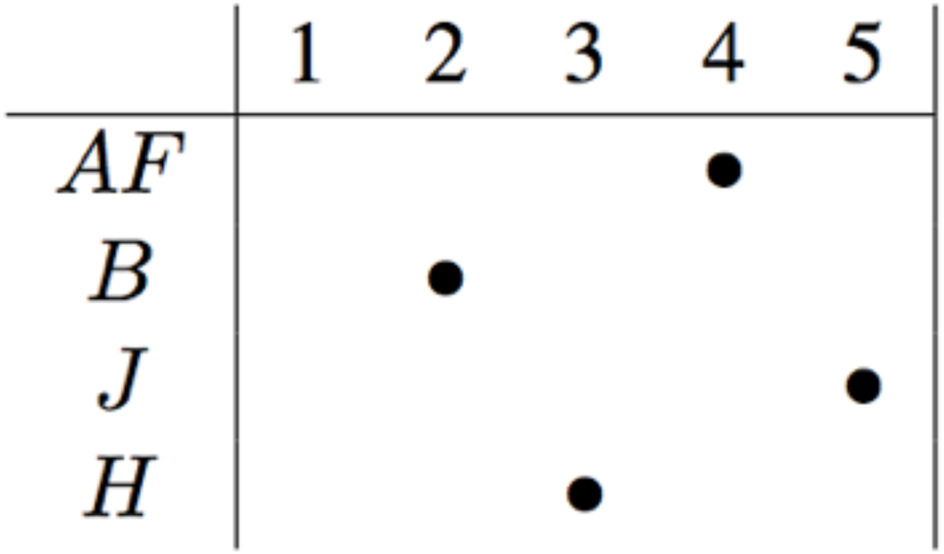
- Then, distribute rankings across systems:

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:

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

$$B < F, B < H, B < J$$

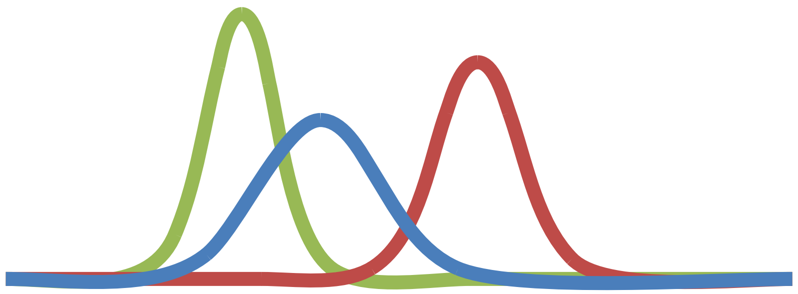
$$F > H, F < J$$

$$H < J$$

→ System Ranking

- Pairwise sentence rankings are aggregated and used to compute the system ranking

Herbrich et al. (2006)

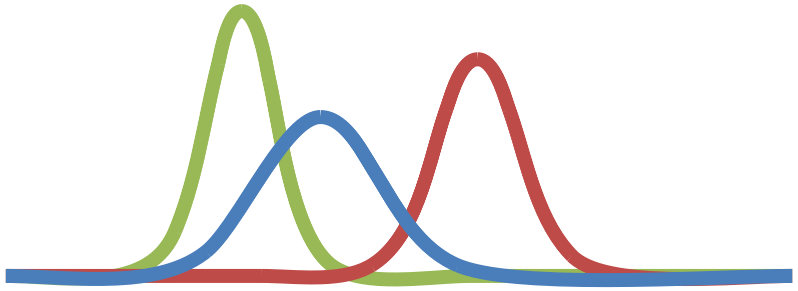


Hopkins & May (2013), Sakaguchi et al. (2014)

→ System Ranking

- Pairwise sentence rankings are aggregated and used to compute the system ranking

- As with WMT14, we used TrueSkill *Herbrich et al. (2006)*
  - Online method, maintains a Gaussian for each system



  - Updates means as games are played
  - Updates proportional to the outcome surprisal

Hopkins & May (2013), Sakaguchi et al. (2014)

Clustering

- A total system ranking is somewhat bogus
  - Lots of similar approaches, same underlying tech
  - Cycles present (Lopez, WMT 2012)
- Instead, compute partial orders, or clusters:
  - Compute rank of each system over 1,000 bootstrap-resampled folds
  - Throw out top and bottom 25 ranks, collect ranges
  - Groups systems by non-overlapping ranges

Koehn (IWSLT 2013)

Participation

- 68 entries from 24 institutions
- +7 anonymized commercial, online, and rule-based systems
    - New! Finnish

Participation

- 68 entries from 24 institutions
- +7 anonymized commercial, online, and rule-based systems
  - New! Finnish

KIT
Karlsruhe Institute of Technology

AVANCEZ
1829

TUM
TECHNISCHE
ITÄT
EN

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

THE UNIVERSITY OF EDINBURGH

AIR FORCE RESEARCH LABORATORY

RWTH AACHEN UNIVERSITY

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

DFKI
Deutsch
Forschu
für Künstliche
Intelligenz GmbH

CNRS LIMSI

Universität
Stuttgart

A?
Aalto University

Limsi

Я

PROMT®

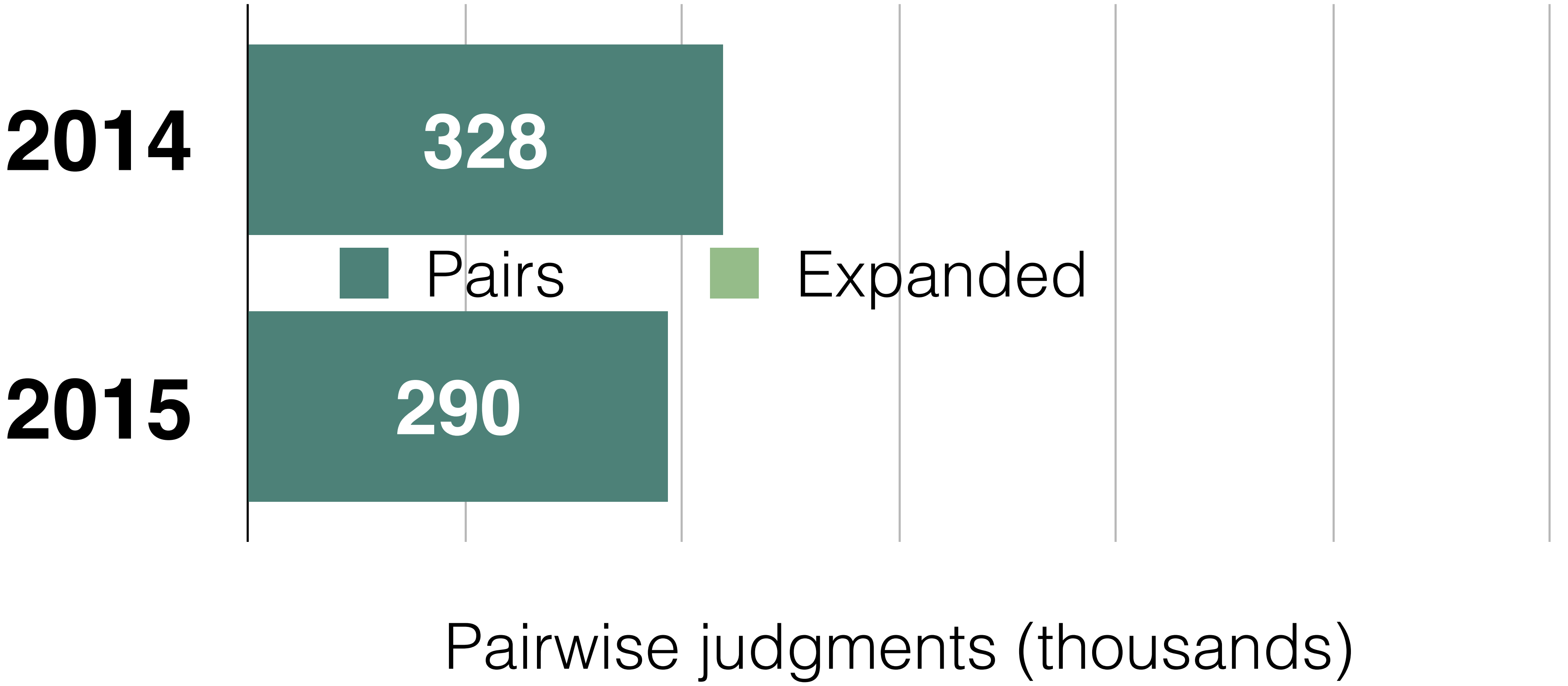
Carnegie
Mellon
University

Université
de Montréal

PROMT®
TRANSLATOR

Data collected

- 137 trusted annotators

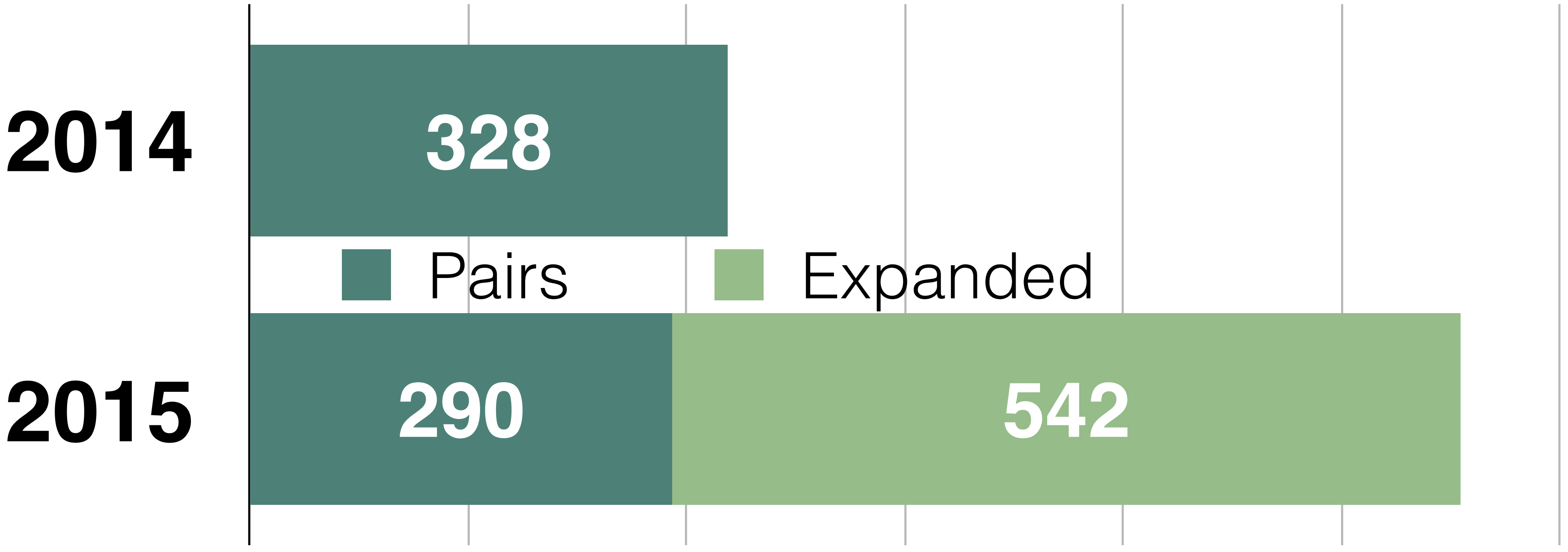


- Punctuation was ignored in collapsing

statmt.org/wmt15/results.html

Data collected

- 137 trusted annotators

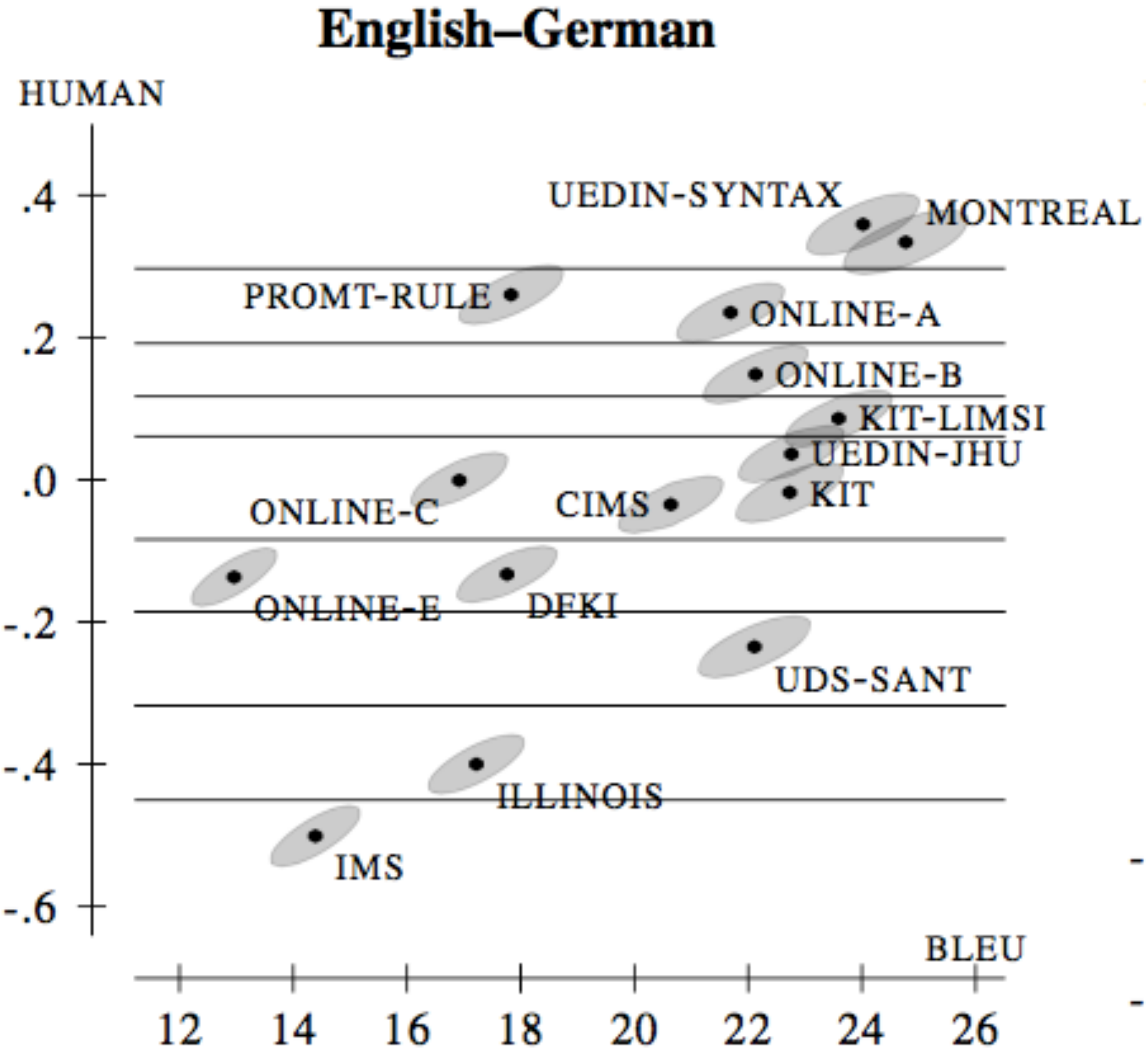


Pairwise judgments (thousands)

- Punctuation was ignored in collapsing

statmt.org/wmt15/results.html

Comparison with BLEU



Results

Czech–English

cluster	constrained	not constrained
1		online-B
2	uedin-jhu	
3	uedin-syntax, montreal	
4		online-A
5	cu-tecto	
6	tt-bleu-mira-d, tt-illc-uva, tt-bleu-mert, tt-afrl, tt-usaar-tuna	
7	tt-dcu, tt-meteor-cmu, tt-bleu-mira-sp, tt-hkust-meant, illinois	

English–Czech

cluster	constrained	not constrained
1		cu-chimera
2	uedin-jhu	online-b
3	montreal	
4		online-a
5	uedin-syntax	
6	cu-tecto	
7		commercial1
8	tt-dcu, tt-afrl, tt-bleu-mira-d	
9	tt-usaar-tuna	
10	tt-bleu-mert	
11	tt-meteor-cmu	
12	tt-bleu-mira-sp	

Russian–English

cluster	constrained	not constrained
1		online-g
2		online-b
3	afrl-mit-pb, afrl-mit-fac, afrl-mit-h, limsi-ncode, uedin-syntax, uedin-jhu	promt-rule, online-a
4	usaar-gacha	
5	usaar-gacha	
6		online-f

English–Russian

cluster	constrained	not constrained
1		promt-rule
2		online-g
3		online-b
4	limsi-ncode	online-a
5	uedin-jhu	
6	uedin-syntax	
7	usaar-gacha	
8	usaar-gacha	
9		online-f

German–English

cluster	constrained	not constrained
1		online-b
2	uedin-jhu, uedin-syntax, kit	online-a
3	rwth, montreal	
4	illinois	dfki, online-c
5		online-f
6	macau	online-e

English–German

cluster	constrained	not constrained
1	uedin-syntax, montreal	
2		prompt-rule, online-a
3		online-b
4	kit-limsi	
5	uedin-jhu, kit, cims	online-f, online-c
6		dfki, online-e
7	uds-sant	
8	illinois	
9	ims	

French–English

cluster	constrained	not constrained
1	limsi-cnrs, uedin-jhu	online-b
2	macau	online-a
3		online-f
4		online-e

English–French

cluster	constrained	not constrained
1	limsi-cnrs	
2	uedin-jhu	online-a, online-b
3	cims	
4		online-f
5		online-e

Finnish–English

cluster	constrained	not constrained
1		online-b
2	abumatran-comb, uedin-syntax, illinois	promt-smt, online-a, uu, uedin-jhu
3	abumatran-hfs	
4	montreal	
5	abumatran	
6	sheff-stem	limsi, sheffield

English–Finnish

cluster	constrained	not constrained
1		online-b
2		online-a
3		uu
4		abumatran-comb
5	abumatran-comb	
6	aalta, uedin-syntax	abumatran
7	cmu	
8	chalmers	

Looking forward

Looking forward

- Pilot: return to direct evaluation (Graham et al., 2015)

Looking forward

- Pilot: return to direct evaluation (Graham et al., 2015)
- Potential advantages:
  - Direct measure of the pursued quality
  - Conceptually simpler?
  - O(n) instead of O(n^2)
  - More statistically significant pairwise cmps.