Simulating Human Judgment in Machine Translation Evaluation Campaigns

Philipp Koehn, University of Edinburgh

December 7, 2012

Manual Evaluation of Machine Translation

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Main Questions

- Goal: Statement about relative quality of systems
- How to rank systems?
- Confidence bounds for rankings?
- How many judgments needed?

Related Work

- Pairwise ranking common practice in research papers
- Obtain ranking of multiple systems based on pairwise rankings
  - rank by ratio of wins vs. losses, ignoring ties
    [Bojar et al., WMT2011]
  - minimize number of pairwise ranking violations
    [Lopez, WMT2012]
  - double seeded knockout with consolation tournament
    [Federico et al., IWSLT2012]
- HyTER [Dreyer and Marcu, NAACL2012]

Idea

- **Problem**

  Hard to assess manual evaluation methods

  — there is no gold standard!

- **Solution**

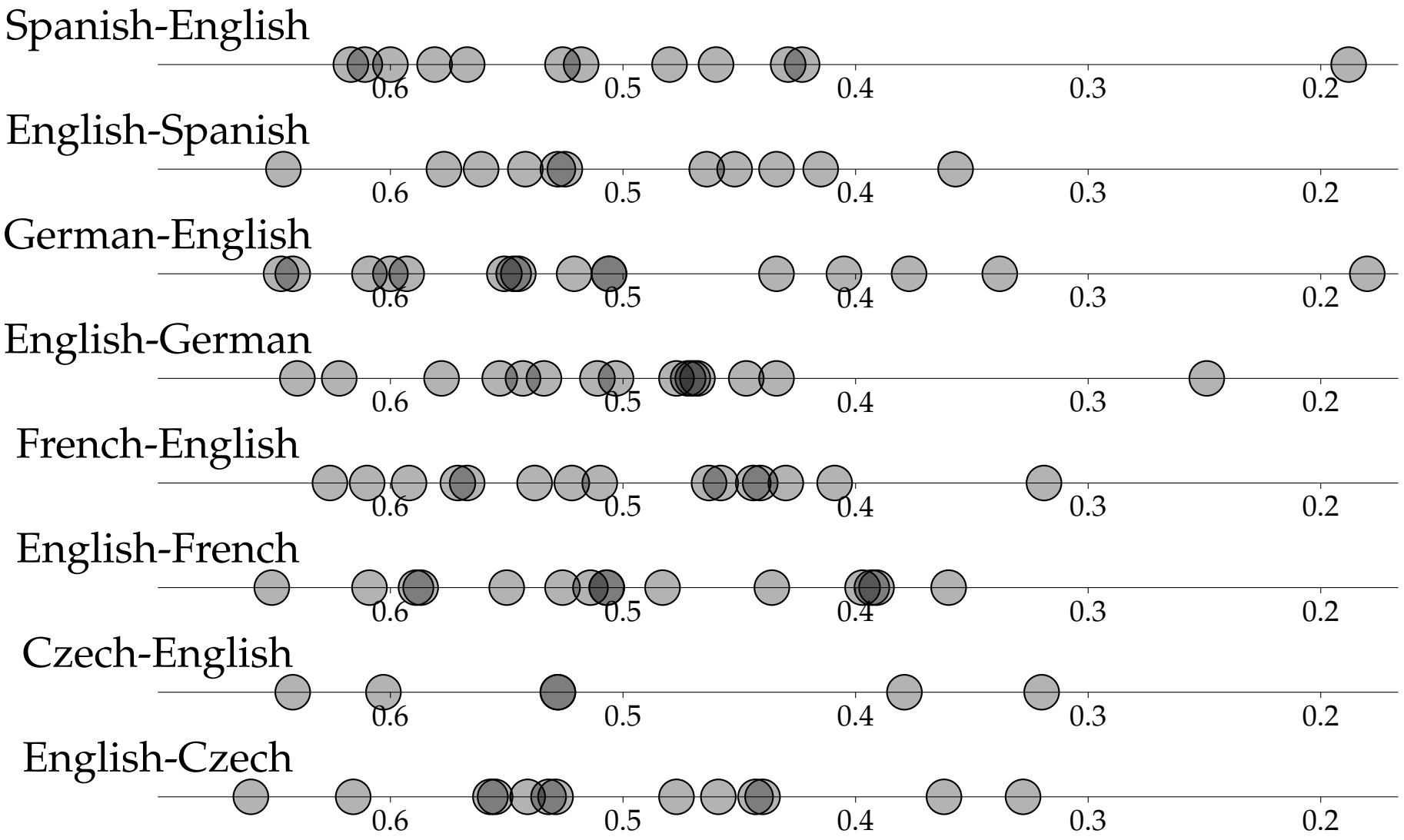
  Simulation of human judgments

Setup

Definitions

- n **systems** $S = \{S_1, ...S_n\}$
- each system average **quality** μ_n
- evaluation **experiment** $E = (\mu_1, ...\mu_n)$:
  each μ_n drawn from uniform distribution [0;10]

Average Ratio of Wins



Sampling Judgments

For each evaluation experiment E draw sample of judgments J_E, by repeating:

- randomly select sets of 5 systems $F_{E,i} = \{s_a, s_b, s_c, s_d, s_e\}$
- each system $j \in F_{E,i}$ produces a translation with a translation quality $q_{E,i,j}$ from normal distribution $\mathcal{N}(\mu_j, \sigma^2)$
- extract set of 10 $(= \frac{5 \times 4}{2})$ pairwise rankings $\{(j_1, j_2) | q_{E,i,j_1} > q_{E,i,j_2}\}$

Remarks

- Range of the average quality interval [0;10] is chosen arbitrarily
- Normal distribution of systems roughly matches WMT systems
- Variance σ^2 same for all systems
- Added complexity because of comparing 5 systems at once
- Ignoring ties
- Not addressing issue of *perceived* translation quality by judge

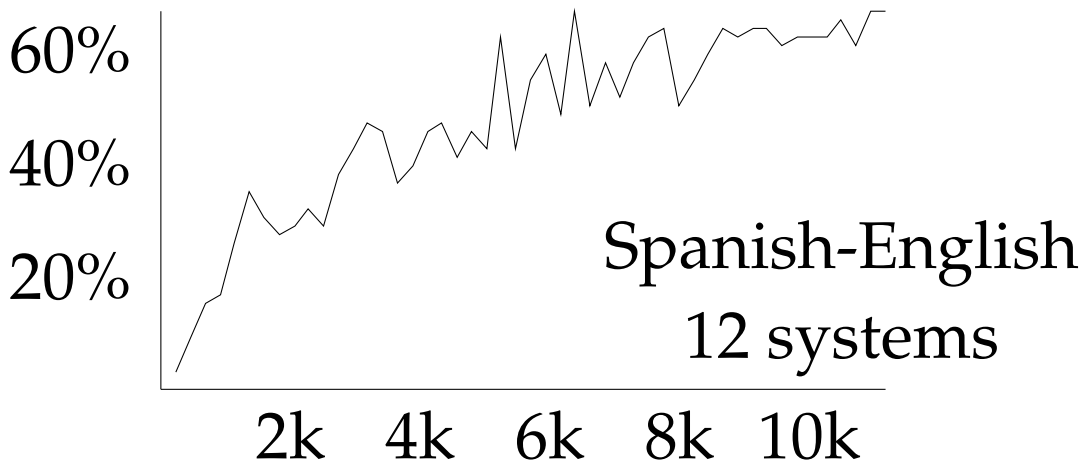
Head to Head Comparison for Czech-English

- For each pair of systems
  - compute ratio of wins
  - apply sign test to assess statistical significance

	CU-BOJAR	JHU	ONLINE-A	ONLINE-B	UEDIN	UK
CU-BOJAR	–	.29⋆	.43	**.53**⋆	**.47**⋆	.31⋆
JHU	**.59**⋆	–	**.59**⋆	**.67**⋆	**.65**⋆	**.44**⋆
ONLINE-A	**.44**	.28⋆	–	**.52**⋆	**.46**⋆	.32⋆
ONLINE-B	.36⋆	.23⋆	.34⋆	–	.38⋆	.25⋆
UEDIN	.36⋆	.23⋆	.36⋆	**.48**⋆	–	.27⋆
UK	**.56**⋆	.33⋆	**.56**⋆	**.63**⋆	**.60**⋆	–
> others	0.53	0.32	0.53	**0.65**	0.60	0.37

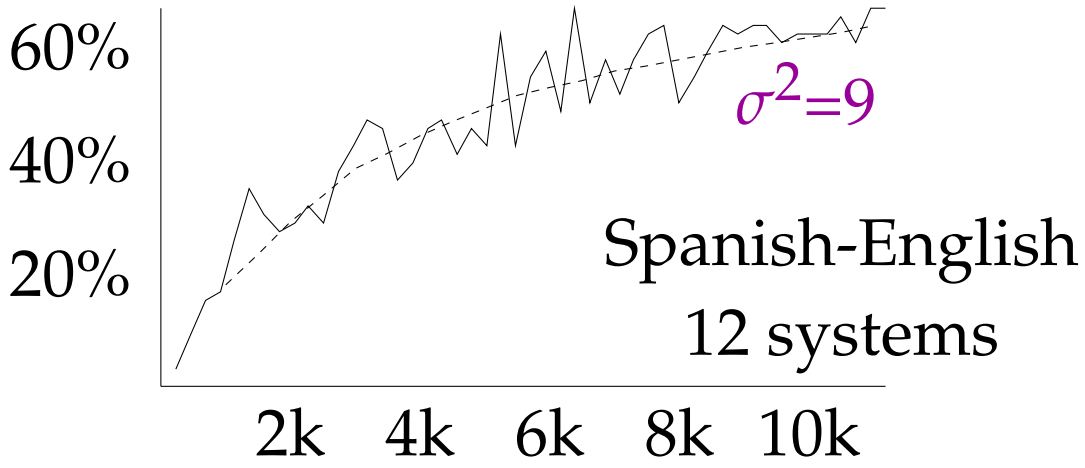
Distinguished Systems

The more judgments
the more system pairs statistically significant difference
according to sign test (p=0.05)



Distinguished Systems

The more judgments
the more system pairs statistically significant difference
according to sign test (p=0.05)



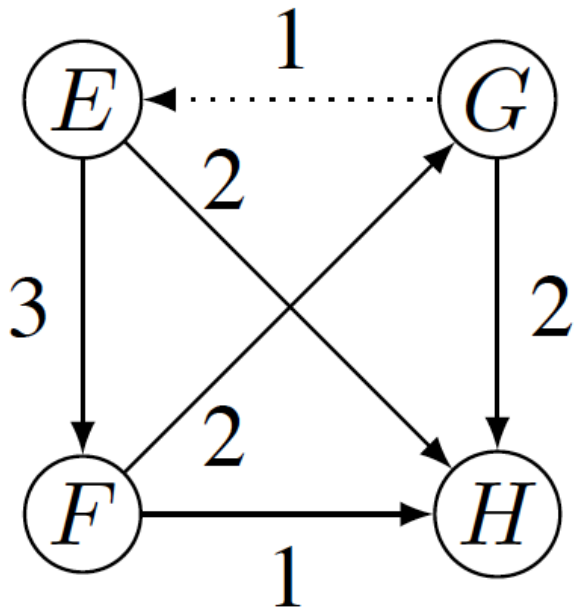
Compared to simulation (average over 1000 experiments)

Ranking Methods

- Task: given pairwise rankings, obtain overall system ranking
- Three methods
  - Ranking violations [Lopez, WMT2012]
  - Win ratio [Bojar et al., WMT2011]
  - Expected win ratio [Callison-Burch et al., WMT2012]
- Evaluation: ranking error
  (i.e., bad system ranked above good system)

Lopez, 2012

- Given graph of pairwise system distinctions (wins minus losses):



- Find ranking with minimum number of violations
- Here: E F G H (1 error, E→G)

Bojar, 2011 / Expected Wins

- Bojar et al., 2011

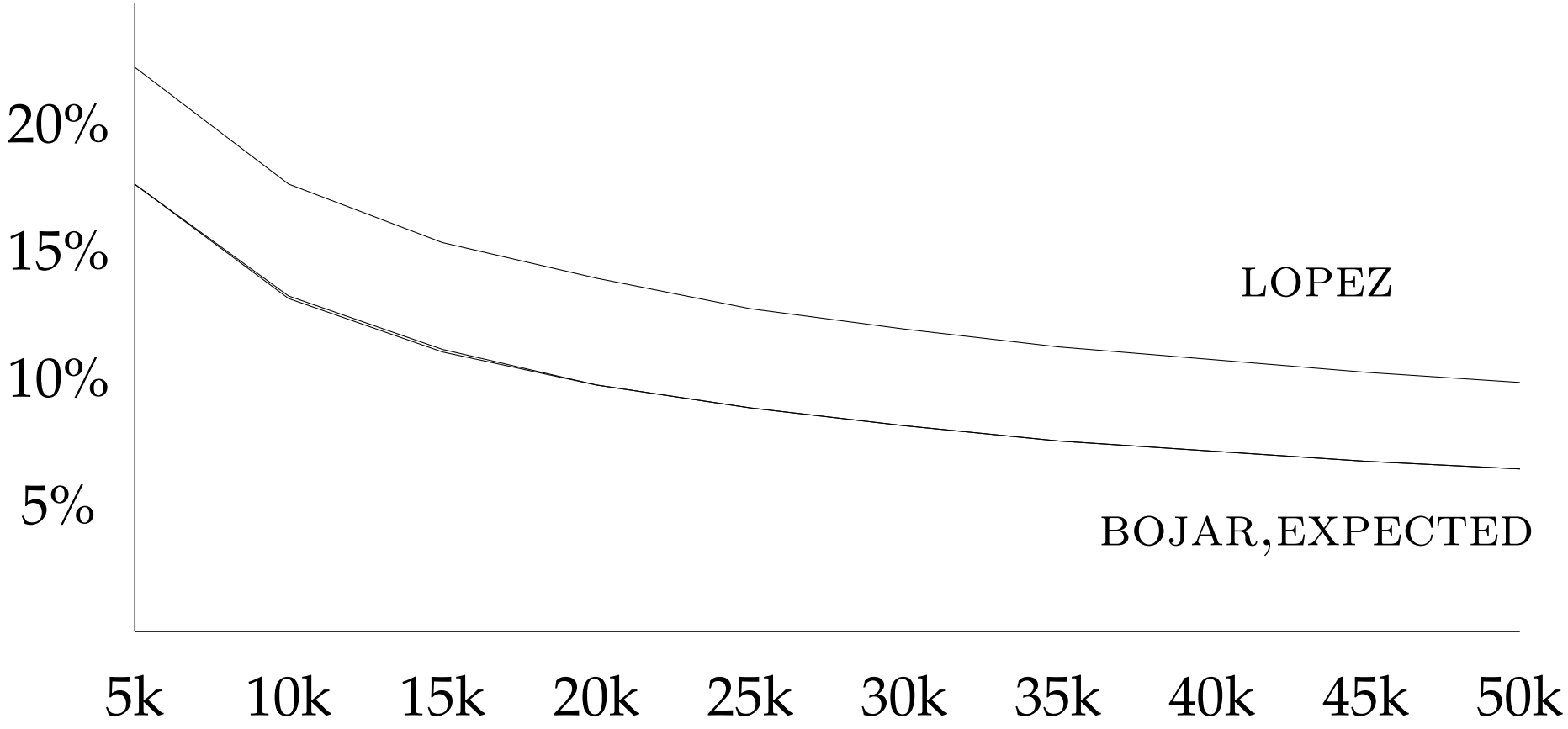
$$\text{score}(S_j) = \frac{\sum_{k,k \neq j} \text{win}(S_j, S_k)}{\sum_{k,k \neq j} \text{win}(S_j, S_k) + \text{loss}(S_j, S_k)}$$

- Expected wins [WMT 2012]

$$\text{score}(S_j) = \frac{1}{n} \sum_{k,k \neq j} \frac{\text{win}(S_j, S_k)}{\text{win}(S_j, S_k) + \text{loss}(S_j, S_k)}$$

Results

15 systems, $\sigma^2 = 10$



Confidence Bounds

- Two forms of presenting confidence
  - **rank range**

    true rank falls within range
  - **cluster**

    no system in cluster worse than any system in lower cluster

- Two methods to obtain confidences
  - based on pairwise statistically significant distinctions
  - bootstrap resampling

Example: English-Czech WMT 2012

Rank	Range	Expected Wins	System
1	1	0.660	CU-DEPFIX
2	2	0.616	ONLINE-B
3	3–6	0.557	UEDIN
4	3–6	0.555	CU-TAMCH
5	3–7	0.541	CU-BOJAR
6	4–7	0.532	CU-TECTOMT
7	4–7	0.529	ONLINE-A
8	8–10	0.477	COMMERCIAL1
9	8–11	0.459	COMMERCIAL2
10	9–11	0.443	CU-POOR-COMB
11	9–11	0.440	UK
12	12	0.362	SFU
13	12	0.328	JHU

Extend from Pairwise Distinctions

- Ranges
  - better than 9 systems, worse than 2, indistinguishable from 3

  → rank range 3–6 (out of 15)

- Clusters
  - grouping systems with overlapping rank ranges

$$\forall S_j \exists C_j : S_j \in C_j$$

$$S_j \in C_j, S_j \in C_k \rightarrow C_j = C_k$$

$$C_j \neq C_k \rightarrow \forall S_j \in C_j, S_k \in C_k :$$

$$\text{start}(S_j) > \text{end}(S_k) \text{ or } \text{start}(S_k) > \text{end}(S_j)$$

Bootstrap Resampling

- Given a fixed set of judgments J_E
- Repeat 1000 times
  - sample pairwise rankings from this set (with replacement)
  - rank systems by score
  - record rank for each system
- For each system: remove 25 highest and 25 lowest rank
  $\rightarrow$ report remaining interval

Quality of Confidence Bounds

Ranges

15 systems, $\sigma^2 = 10$

Judgments	Pairwise Method		Bootstrap Method	
$\|J_E\|$	range size	violations	range size	violations
10,000	8.1	0.8%	4.6	3.4%
20,000	6.3	0.8%	3.7	2.4%
30,000	5.4	0.7%	3.3	2.3%
40,000	4.9	0.9%	3.0	2.0%
50,000	4.5	0.9%	2.9	2.1%

Quality of Confidence Bounds

Clusters

15 systems, $\sigma^2 = 10$

Judgments	Pairwise Method		Bootstrap Method	
$\|J_E\|$	clusters	violations	clusters	violations
10,000	1.0	0%	1.8	0.5%
20,000	1.1	0%	3.0	0.5%
30,000	1.4	0%	3.9	0.4%
40,000	1.7	0.1%	4.7	0.4%
50,000	2.0	0.1%	5.3	0.7%

Example: French–English WMT 2012

Rank	Range	Expected Wins	System
1	1–3	0.626	LIMSI
2	1–4	0.610	KIT
3	1–5	0.592	ONLINE-A
4	2–6	0.571	CMU
5	3–7	0.567	ONLINE-B
6	5–8	0.538	UEDIN
7	5–8	0.522	LIUM
8	6–9	0.510	RWTH
9	8–12	0.463	RBMT-1
10	9–13	0.458	RBMT-3
11	9–14	0.444	SFU
12	9–14	0.441	UK
13	10–14	0.430	RBMT-4
14	12–14	0.409	JHU
15	15	0.319	ONLINE-C

How Many Judgments are Needed?

- Depends on...
  - number of systems
  - similarity of systems
  - variance of systems
  - noisiness of judgments

- Information distilled in model variable σ^2, typical values 8-12

- Run simulation with increasing number of judgments
  - WMT12 French–English, ($n = 15$, $\sigma^2 = 10$)
  - collected 13,000 judgments, 50% of pairs different
  - increase to 40,000 judgments $\rightarrow$ 70% of pairs different

Results

n	σ^2	Ratio of significant pairs			
		50%	70%	80%	90%
6	8	1k	4k	8k	30k
6	10	2k	5k	10k	45k
6	12	2k	7k	20k	60k
8	8	2k	6k	14k	60k
8	10	3k	8k	20k	90k
8	12	4k	14k	35k	140k
10	8	4k	10k	25k	100k
10	10	5k	16k	40k	150k
10	12	6k	20k	50k	200k
12	8	5k	15k	35k	140k
12	10	7k	25k	60k	250k
12	12	9k	35k	80k	350k
15	8	8k	25k	50k	200k
15	10	12k	40k	80k	350k
15	12	15k	50k	120k	500k

Conclusions

- Introduced a Monte Carlo model for simulation manual evaluation
- Compared different ranking methods
- New methods to obtain confidence bounds
- Estimates how many judgments needed
  $\rightarrow$ for WMT about three times as many judgements needed

questions?