Direct Assessment

Yvette Graham

August 11, 2016

Direct Assessment (DA) I

- Consideration being given to using DA alone for next year

Reasons:

- High correlation between RR and DA
- It seems like we could get good clusters with (conservatively) half the annotation time
- Computed as follows:
  - we require 100 hits per system submission, average 5 min per hit, so 500 minutes = 8 hours
  - DA at 500 translations is what we might need (maybe more in some cases), and that takes about 2.5 hours (half hour per hit)

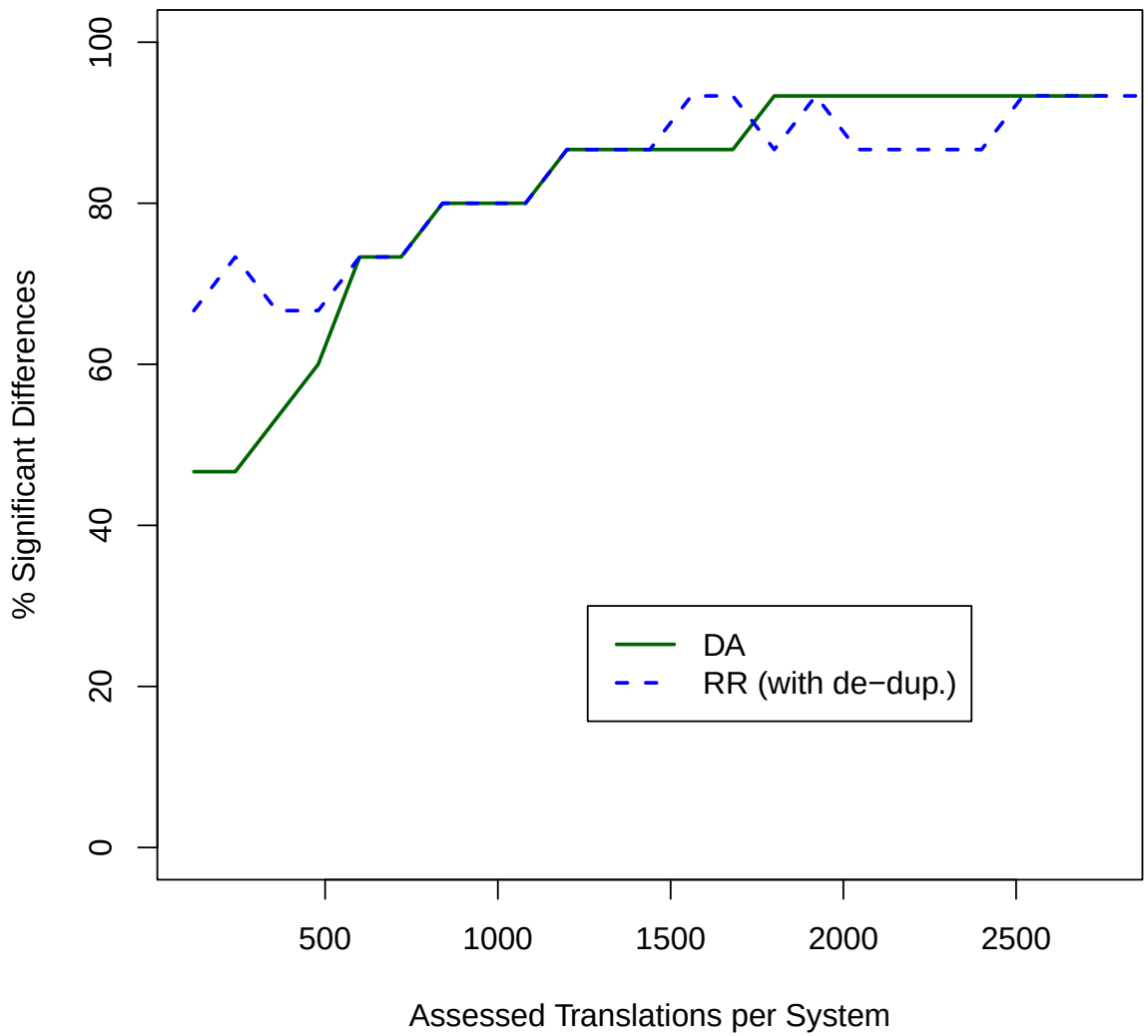
Direct Assessment (DA) II

- English side can be completely crowdsourced
- Leaves researchers responsible only for tasks where we can't find crowdsourced workers

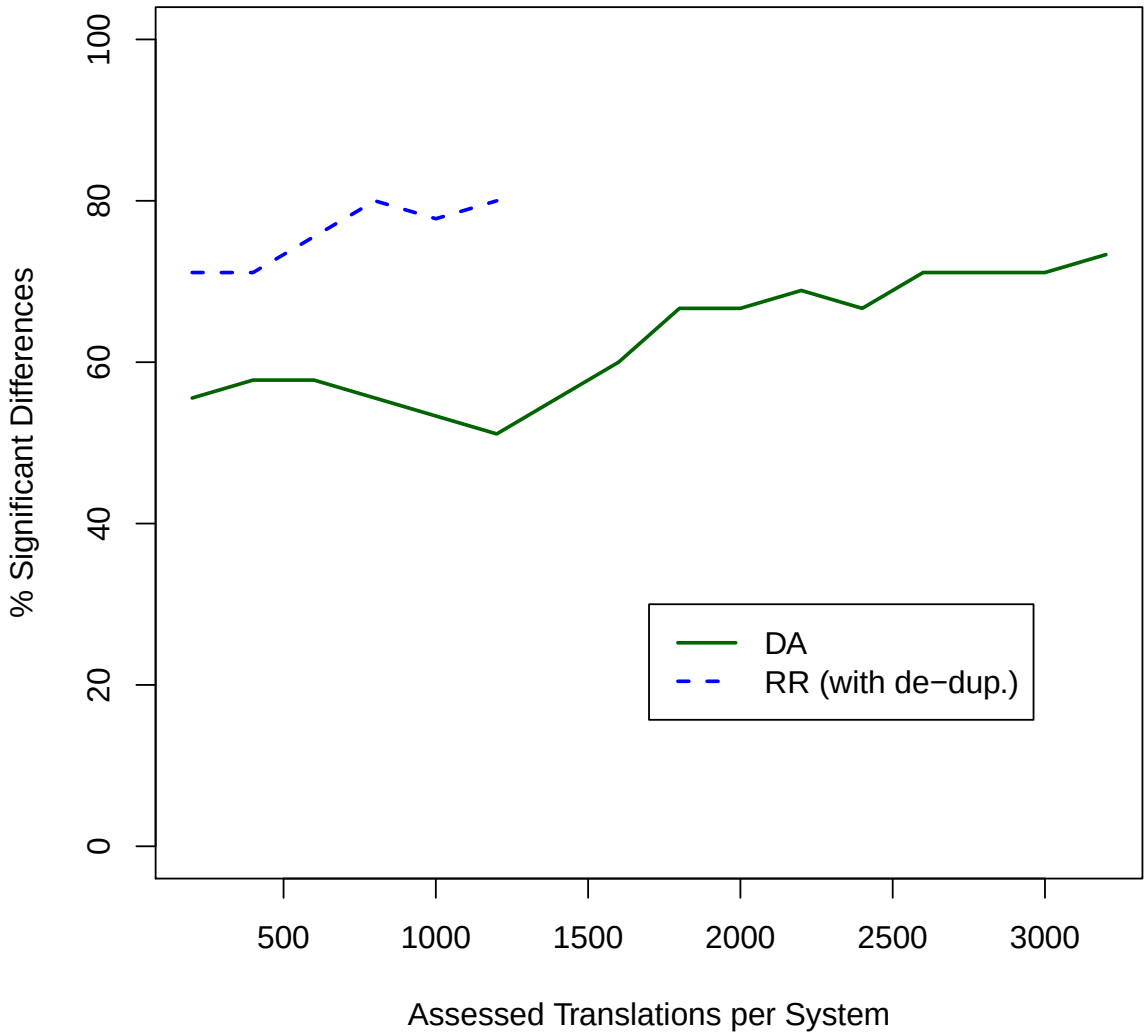
Correlation of RR and DA

cs-en	0.997
fi-en	0.996
tr-en	0.988
de-en	0.964
ru-en	0.961
ro-en	0.920
en-ru	0.975

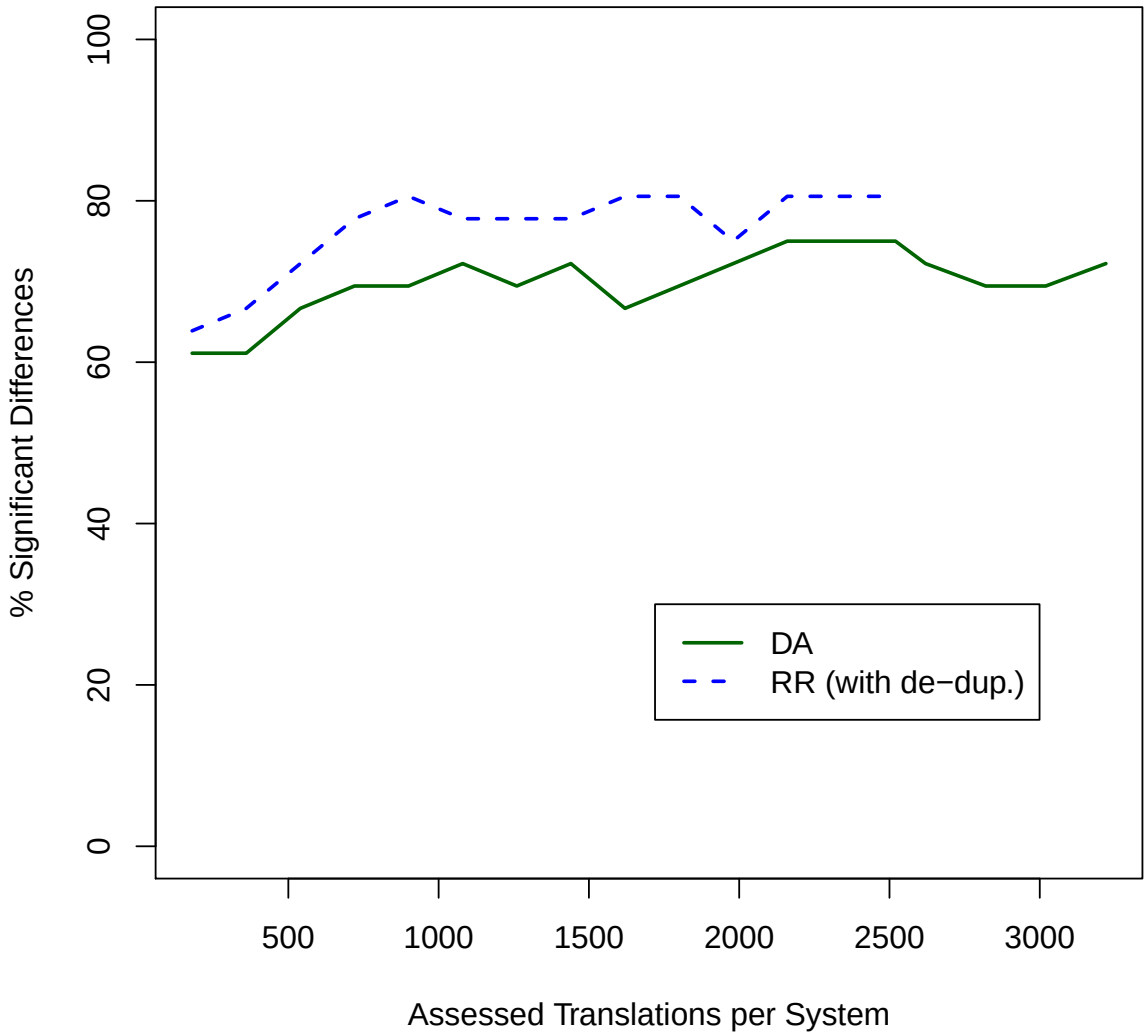
Human Assessment vs Significant Differences: CS-EN



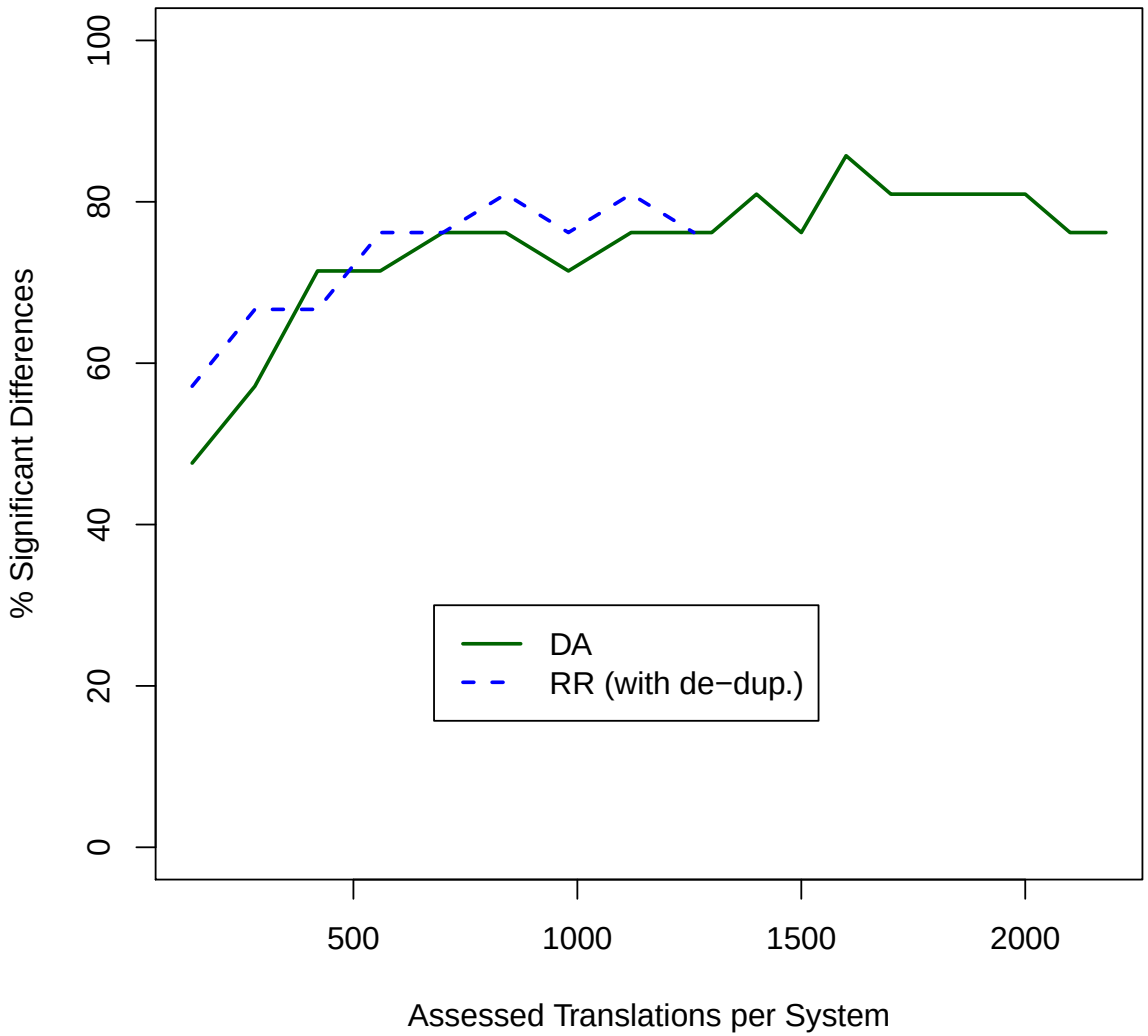
DE-EN


% Significant Differences
0
20
40
60
80
100
500
1000
1500
2000
2500
3000
Assessed Translations per System
DA
RR (with de−dup.)

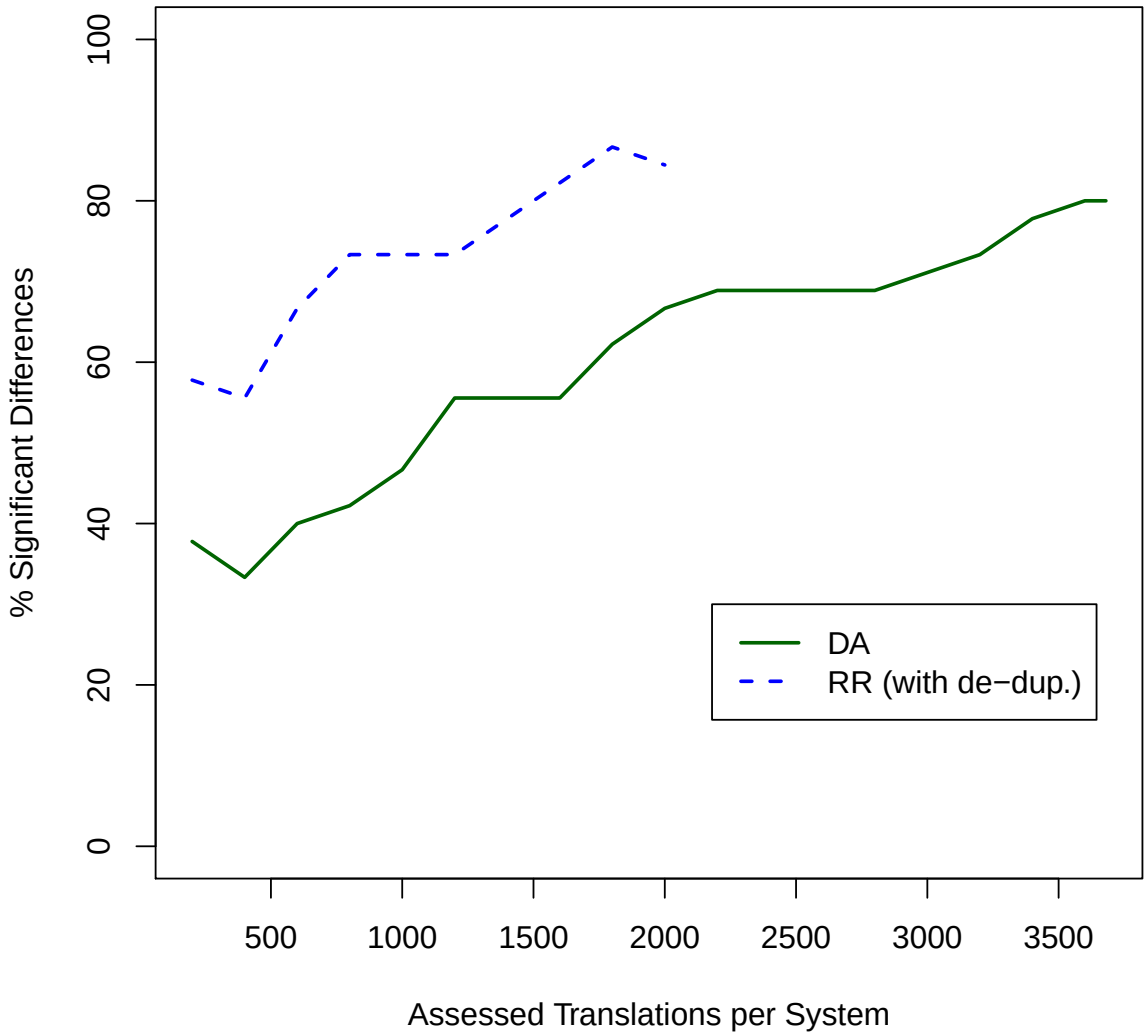
FI-EN


% Significant Differences
0
20
40
60
80
100
500
1000
1500
2000
2500
3000
Assessed Translations per System
DA
RR (with de−dup.)

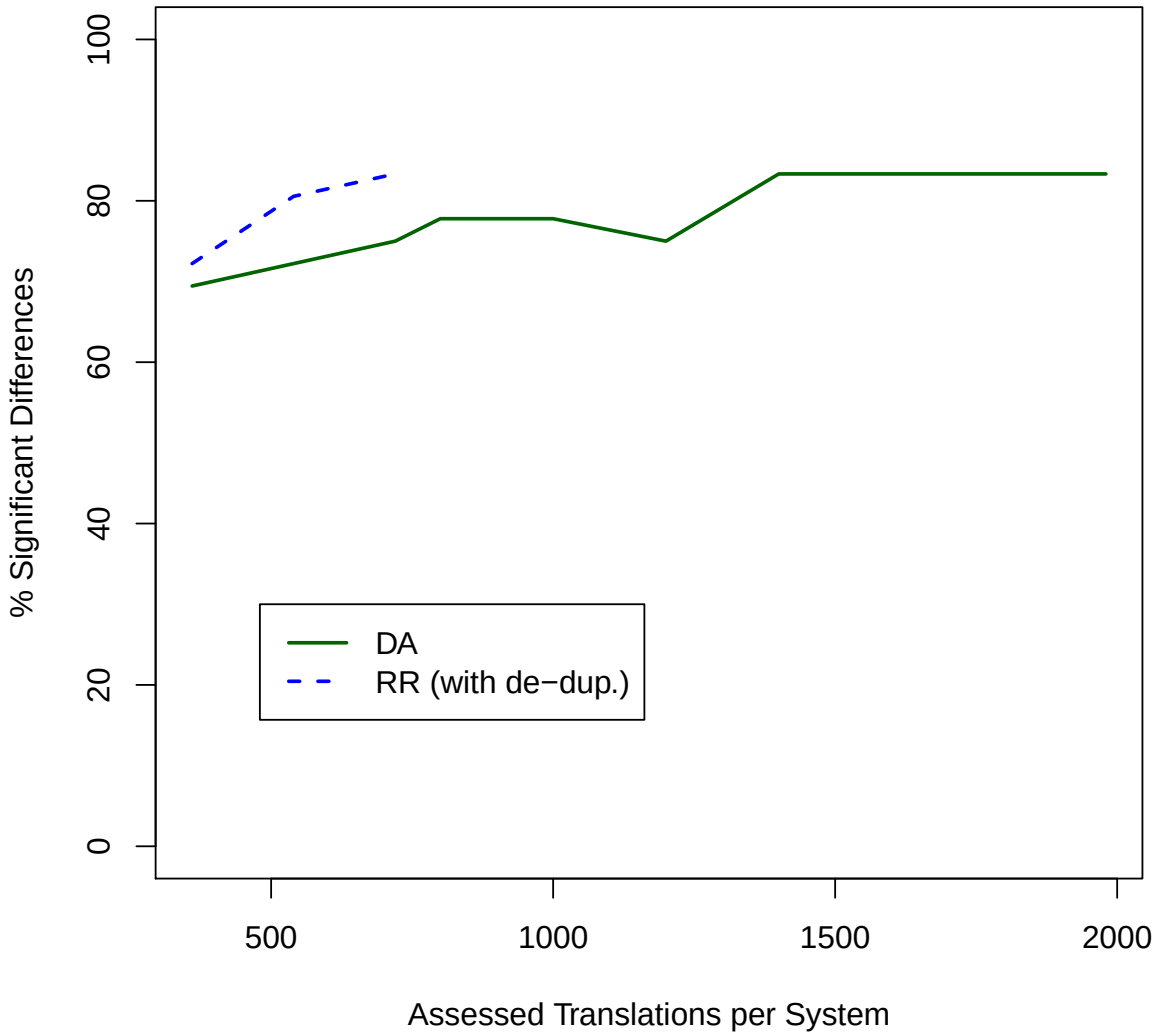
RO-EN


% Significant Differences
0
20
40
60
80
100
500
1000
1500
2000
Assessed Translations per System
DA
RR (with de−dup.)

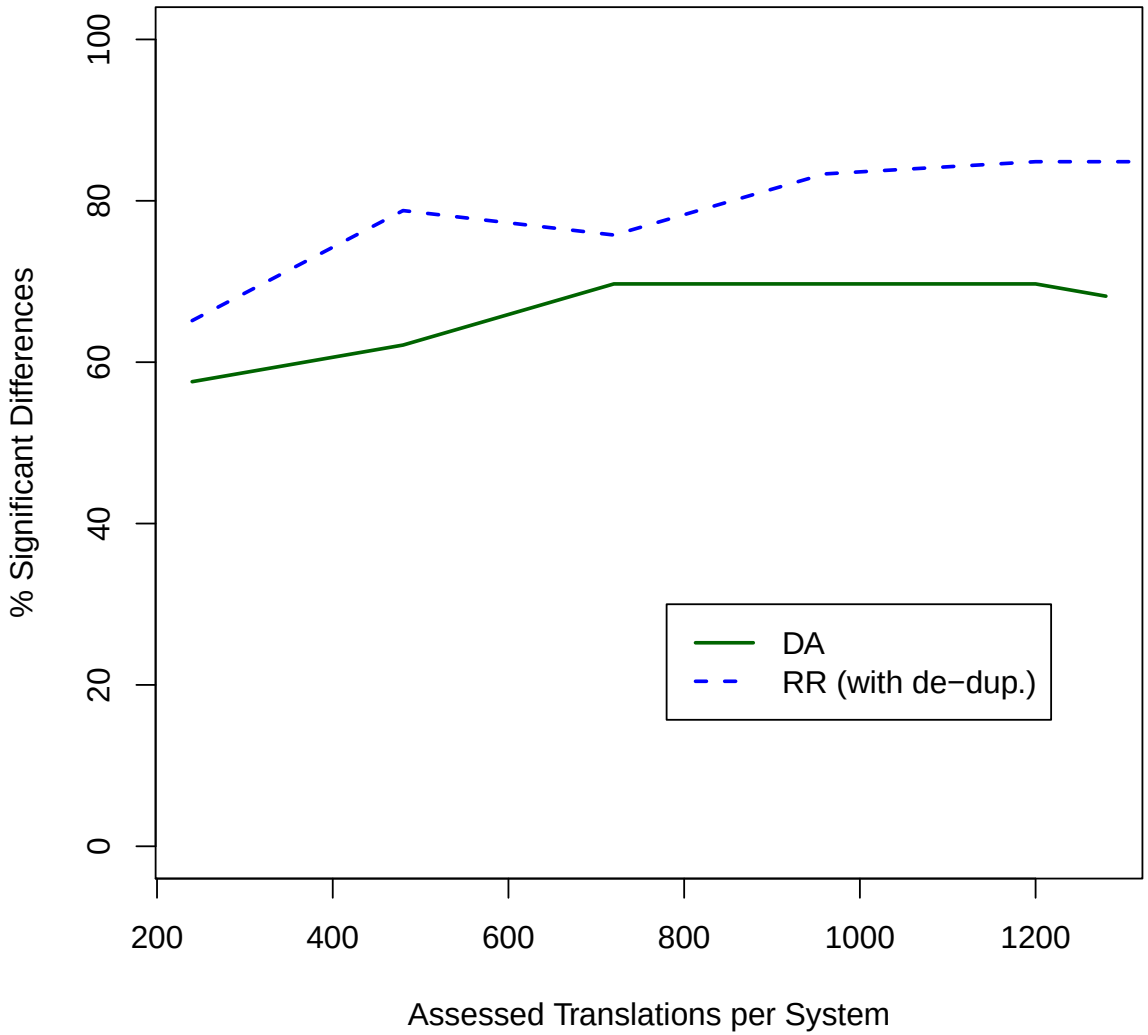
RU-EN


% Significant Differences
0
20
40
60
80
100
500
1000
1500
2000
2500
3000
3500
Assessed Translations per System
DA
RR (with de−dup.)

TR-EN


% Significant Differences
0
20
40
60
80
100
500
1000
1500
2000
Assessed Translations per System
DA
RR (with de−dup.)

EN-RU


% Significant Differences
0
20
40
60
80
100
200
400
600
800
1000
1200
Assessed Translations per System
DA
RR (with de−dup.)

Conclusions

- Trial of DA was successful overall
- No problems crowd-sourcing all to-English language pairs
- Not enough workers for all out-of-English news LPs except English to Russian – those LPs unfortunately must remain the task of participants
- Correlation with RR high across the board
- DA almost achieves as many significant differences as RR but without deduplication

More to come:

- WMT'16 included RR with deduplication and DA without it – Makes the comparison of numbers of judgments difficult
- Future: Compare unexpanded (undeduped) versions to see what effect it had, since this is really an unfair comparison