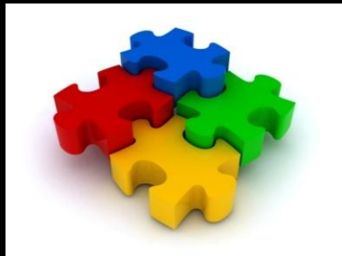
An IR-based Evaluation Framework for Web Search Query Segmentation

Rishiraj Saha Roy and Niloy Ganguly
IIT Kharagpur
India

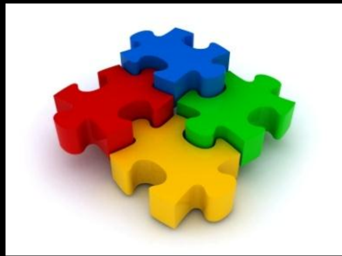
Monojit Choudhury and Srivatsan Laxman
Microsoft Research India
India

ACM SIGIR 2012, Portland
August 15, 2012



Query Segmentation

- Dividing a query into individual semantic units (Bergsma and Wang, 2007)
- Example
  - *history of all saints church south australia* →
  - *history of | all saints church | south australia*
  - *history of all | saints church south | australia*

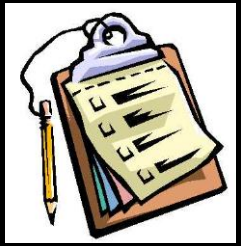


Query Segmentation

- Goes beyond multiword named entity recognition (*gprs config*, *history of*, *how to*)
- Helps in better query understanding
- Can improve IR performance (Bendersky et al. 2009; Li et al. 2011)
- This research: Focus on **evaluation**, not on algorithm

Evaluation till now

- An algorithm segments each query in test set
- A segmented query is matched against the human annotated query using five metrics (Hagen et al. 2011)



Evaluation till now

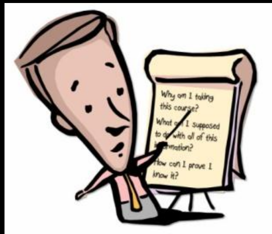
- **Segment Precision** – Fraction of machine segments that match with the human segments
- **Segment Recall** – Fraction of human segments that match with the machine segments
- **Segment F-Score** – Harmonic mean of precision and recall
- **Query Accuracy** – Fraction of queries where machine and human segmentations match exactly
- **Classification Accuracy** – Fraction of boundaries *and* non-boundaries that match between human and machine segmentations

Evaluation till now

- Problems
  - Low inter-annotator agreement on most metrics (≈ 70%) (Tan and Peng 2008)
  - **Human A:** *grand theft auto | san andreas | ps2 | cheats*
  - **Human B:** *grand theft auto san andreas | ps2 cheats*
  - Not clear what should be the guidelines

Evaluation till now

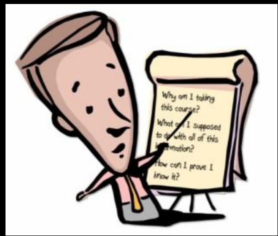
- Problems
  - Humans may not be the best judge as to which segments are best for IR – Humans are not the end users of segmentation!!



Proposed Evaluation Framework

q → **Segmentation Algorithm** → q^s → **IR Engine** → *Ranked list of docs*

- End user of segmentation is the search engine
- An IR performance based evaluation
- **Main challenge:** how to use segmented query for retrieval

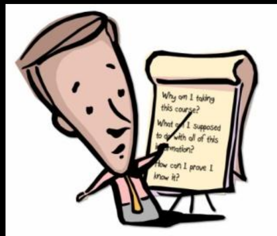


Proposed Evaluation Framework

- Different segments of the same query may need to be matched differently in documents for the best results

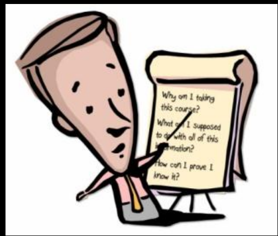
  cannot view* | *word files* | *windows 7

  - Ordered (***windows 7***)
  - Unordered (may have linguistic constraints) (***files** in **word***)
  - Insertions, deletions, transpositions, substitutions (***cannot** properly **view***)
  - MRF models of term dependence (Metzler and Croft, 2005)
- Certain segments need not be matched at all (***view online*, *cheap*, *near***)



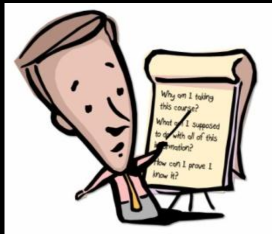
Proposed Evaluation Framework

- Current IR engines do not support these specifications
- Most retrieval systems support use of double quotes (exact match)
- However, simply putting double quotes around all query segments results in very poor retrieval performance!!
- Hagen et al. (2011) explore an evaluation with quotes around all segments, effective only for MWEs and negatively affecting overall results



Proposed Evaluation Framework

- We adopt a less constrained approach
- For each segmentation algorithm output, we generate all quoted versions of segmented query q^s (each segment can be quoted or unquoted)
- 2^k quoted versions for a k-segment query

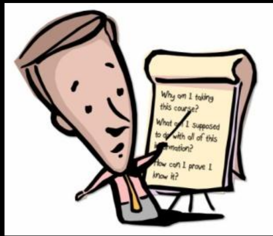


Proposed Evaluation Framework

Segmented query	Quoted versions
	history of all saints church south australia
	history of all saints church "south australia"
	history of "all saints church" south australia
history of | all saints church | south australia	*history of "all saints church" "south australia"*
	"history of" all saints church south australia
	"history of" all saints church "south australia"
	"history of" "all saints church" south australia
	"history of" "all saints church" "south australia"

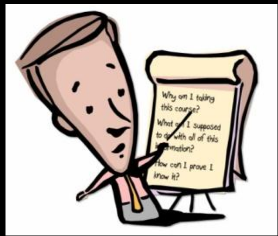
Proposed Evaluation Framework

- Each version issued through IR engine (after query versions are deduplicated)
- IR system retrieves top *k* pages for each quoted version of a query
- Measure performance (eg. nDCG) of each quoted version (using human relevance judgments)



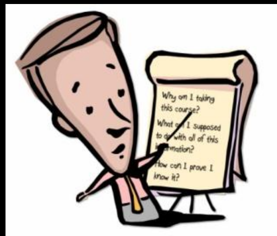
Proposed Evaluation Framework

Segmented query	Quoted versions	Score
	history of all saints church south australia	0.723
	history of all saints church "south australia"	0.788
	history of "all saints church" south australia	0.801
history of | all saints church | south australia	*history of "all saints church" "south australia"*	**0.852**
	"history of" all saints church south australia	0.632
	"history of" all saints church "south australia"	0.645
	"history of" "all saints church" south australia	0.652
	"history of" "all saints church" "south australia"	0.619



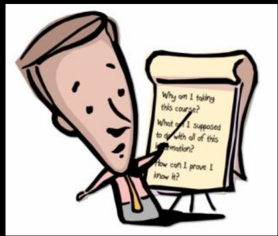
Proposed Evaluation Framework

- **Use of *Oracle*:** *Highest* nDCG from all quoted versions chosen as score achieved by q^s
- Reflects "potential" of a segmented query
- Directly correlates to goodness of segmentation algorithm



Proposed Evaluation Framework

- For each algorithm, compute average oracle *score* over all queries
- **Find gold standard for IR performance:** Also perform *brute force* exhaustive search over all possible quoted versions of a query to find the one with the highest score
- Call it the *best quoted version (BQV (BF))* of a query, irrespective of any segmentation algorithm
  - 2^{n-1} quoted versions for an *n*-word query



Resources Required by Framework

- Any search engine that supports double quotes (Lucene in our experiments)
- Test set of queries
- Document pool
- Query relevance sets (*qrels*): For each query, human relevance judgments for the subset of documents in the pool possibly relevant to the query
- These resources are required for any IR-system evaluation

Dataset

- Query test set
  - 500 test queries (5-8 words) sampled from Bing Australia in May 2010
- Document collection
  - All possible quoted versions of a test query are issued through the Bing API 2.0
  - Top 10 URLs retrieved are deduplicated and added to collection

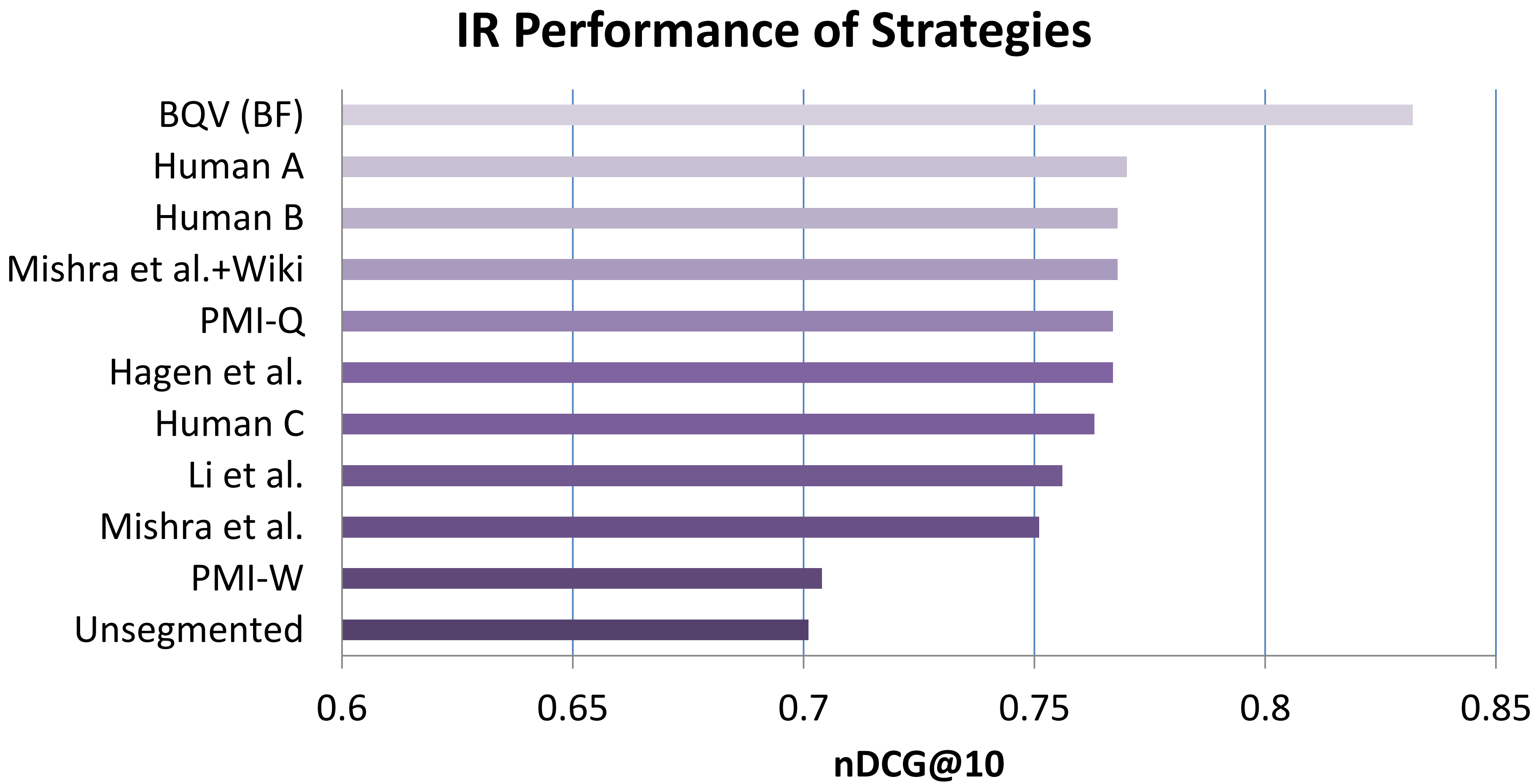
Dataset

- Relevance judgments
  - For each query, three sets of relevance judgments obtained for each URL retrieved for the query
  - Much higher agreement on relevance judgments than human segment boundaries

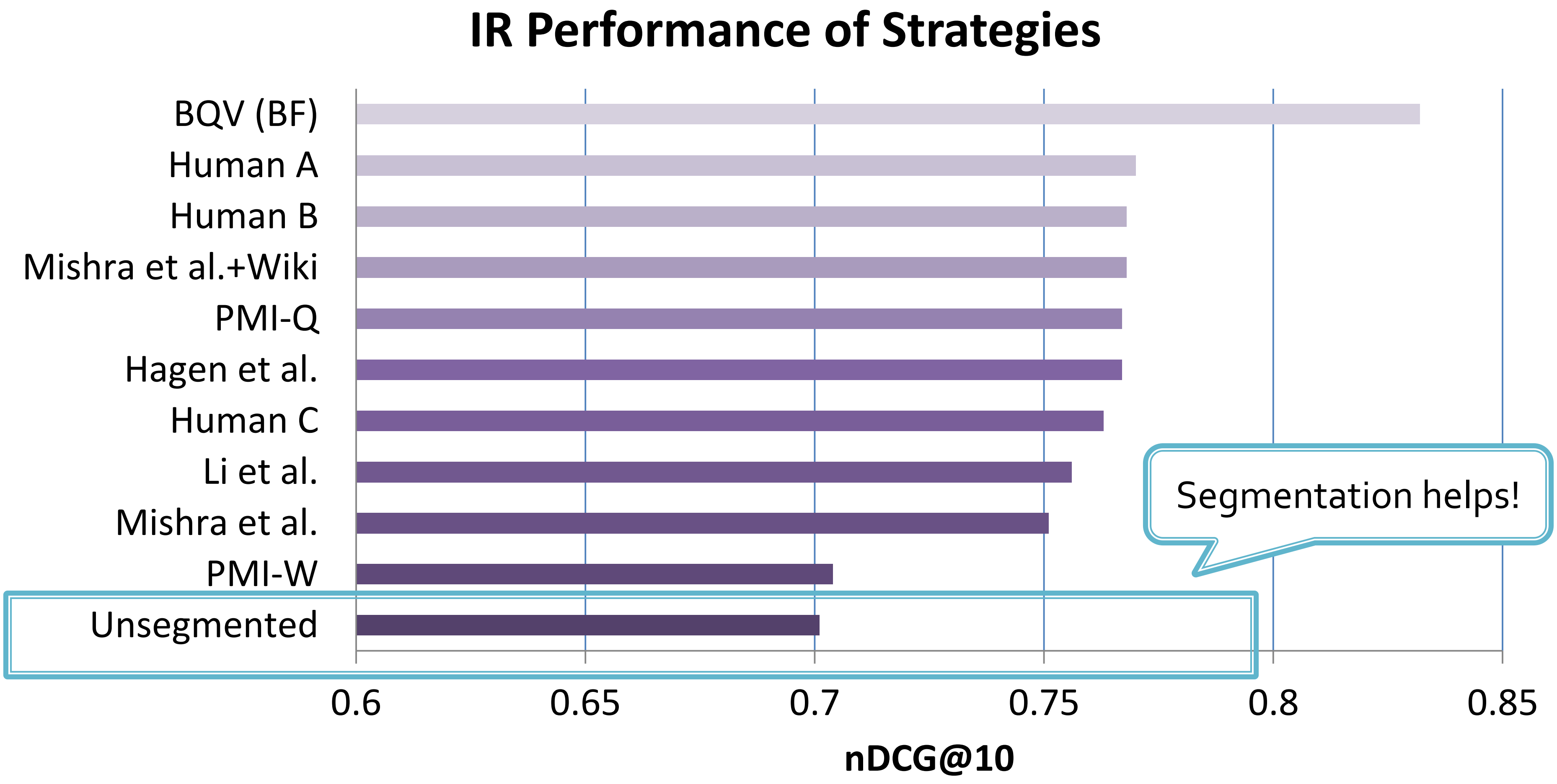
Experiments

- Six segmentation strategies compared on our framework including (four state-of-the-art systems)
  - Li et al. (SIGIR 2011), Hagen et al. (WWW 2011), Mishra et al. (WWW 2011), Mishra et al.+Wiki (SIGIR 2012)
  - Baselines: PMI-W, PMI-Q
- Plus annotations by three human annotators *A*, *B*, *C*

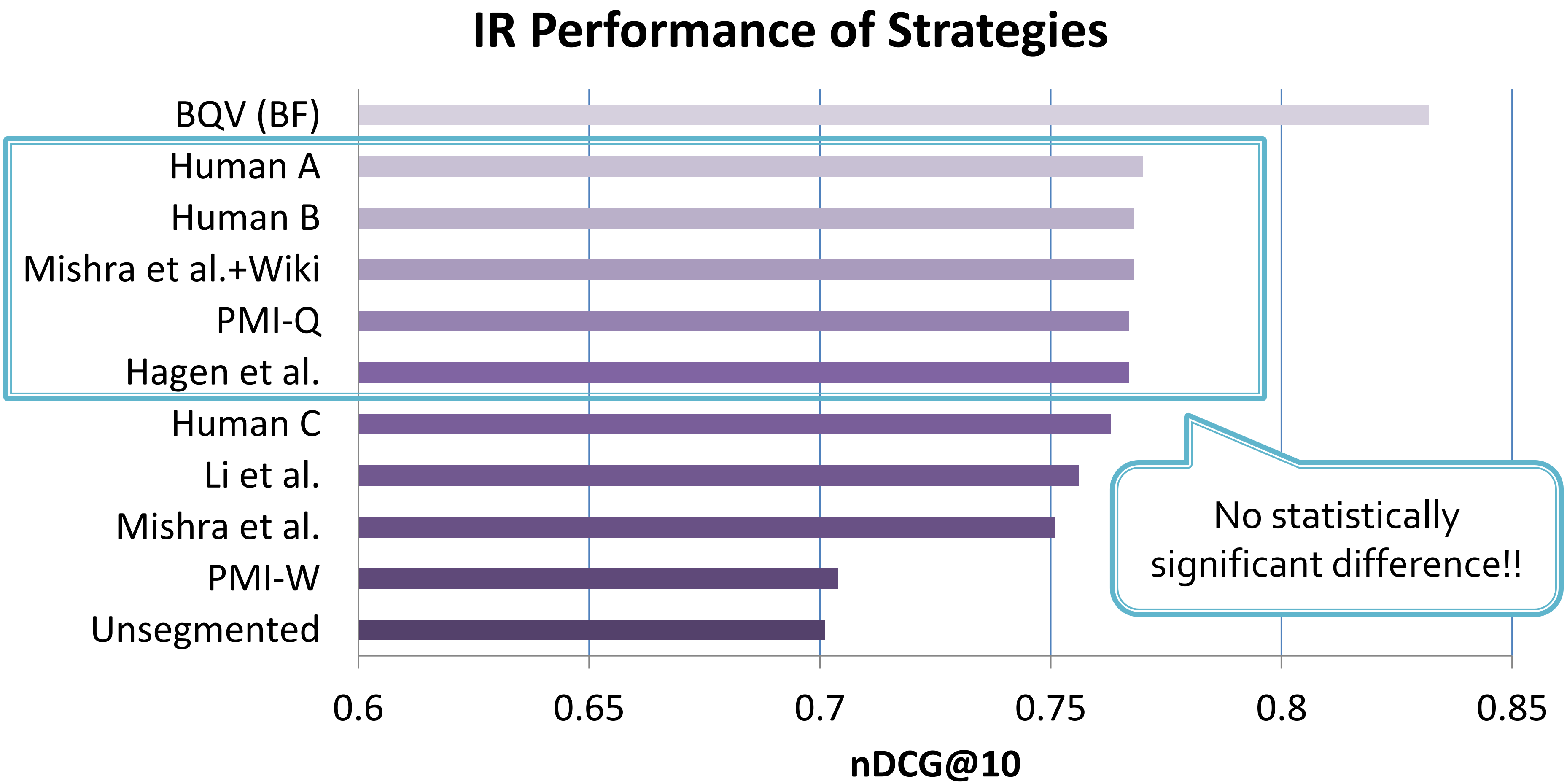
Results


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Results

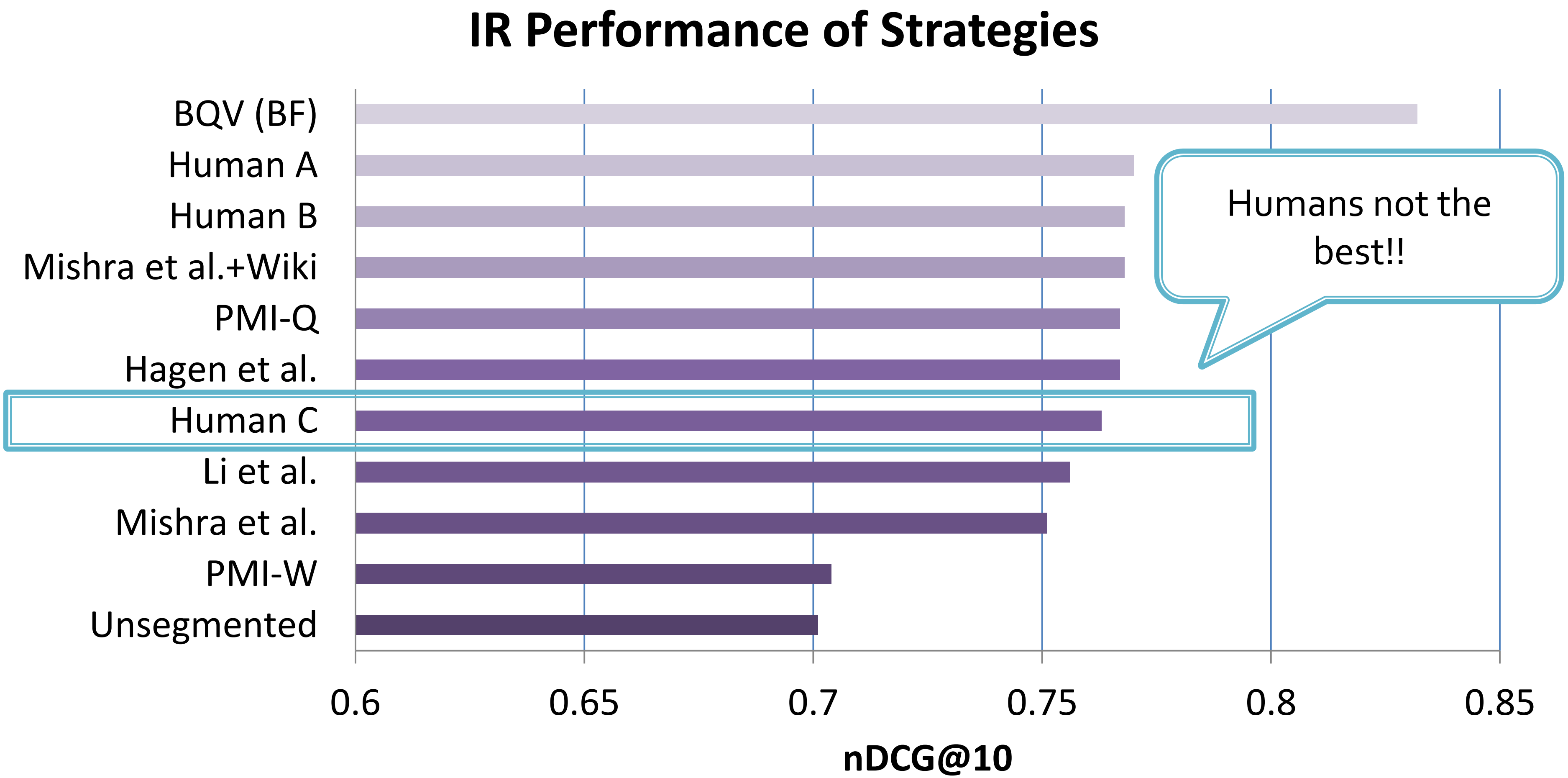

IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
Segmentation helps!
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Results

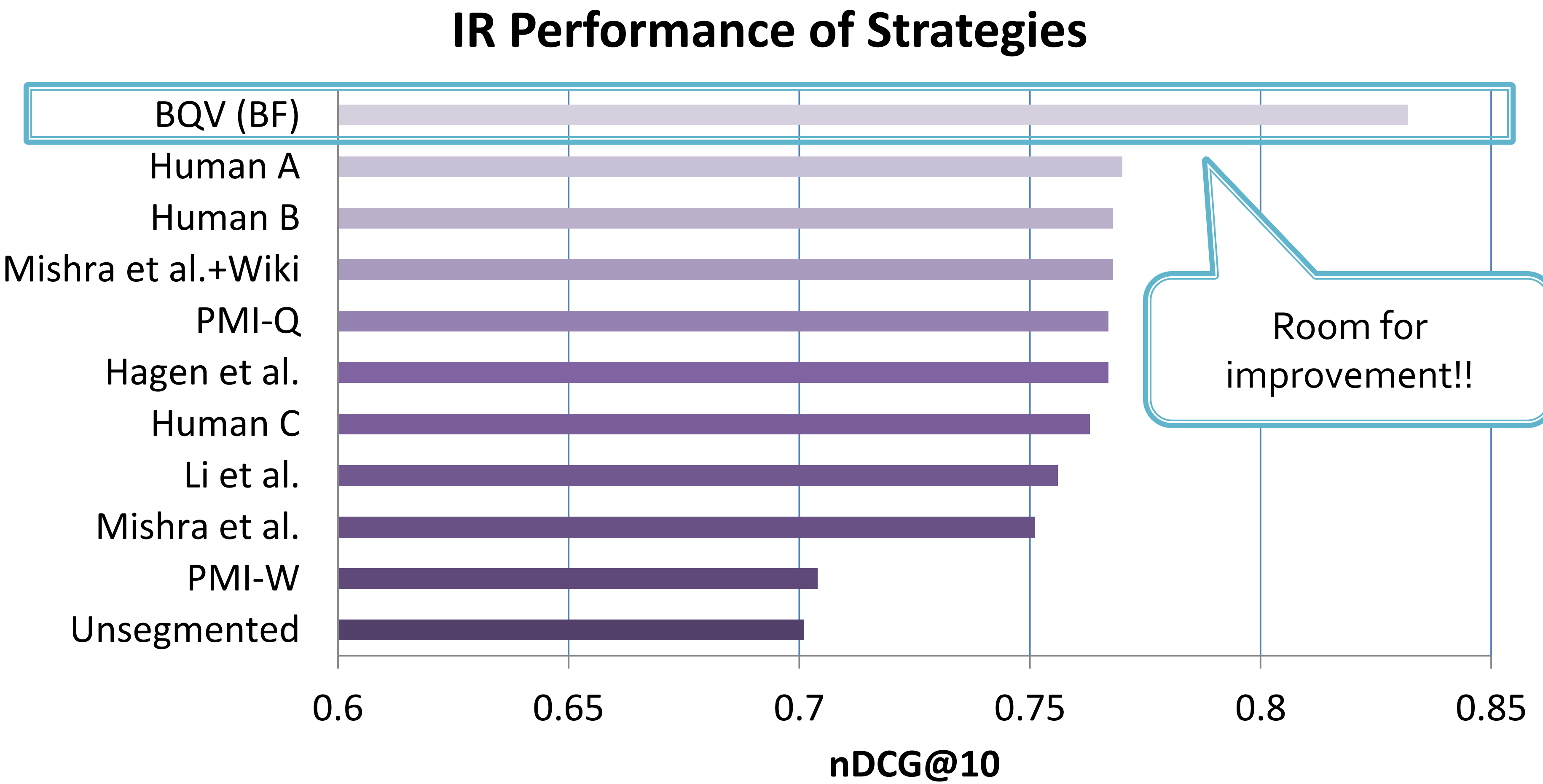


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
No statistically significant difference!!
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Results


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10
Humans not the best!!

Results



IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10
Room for improvement!!

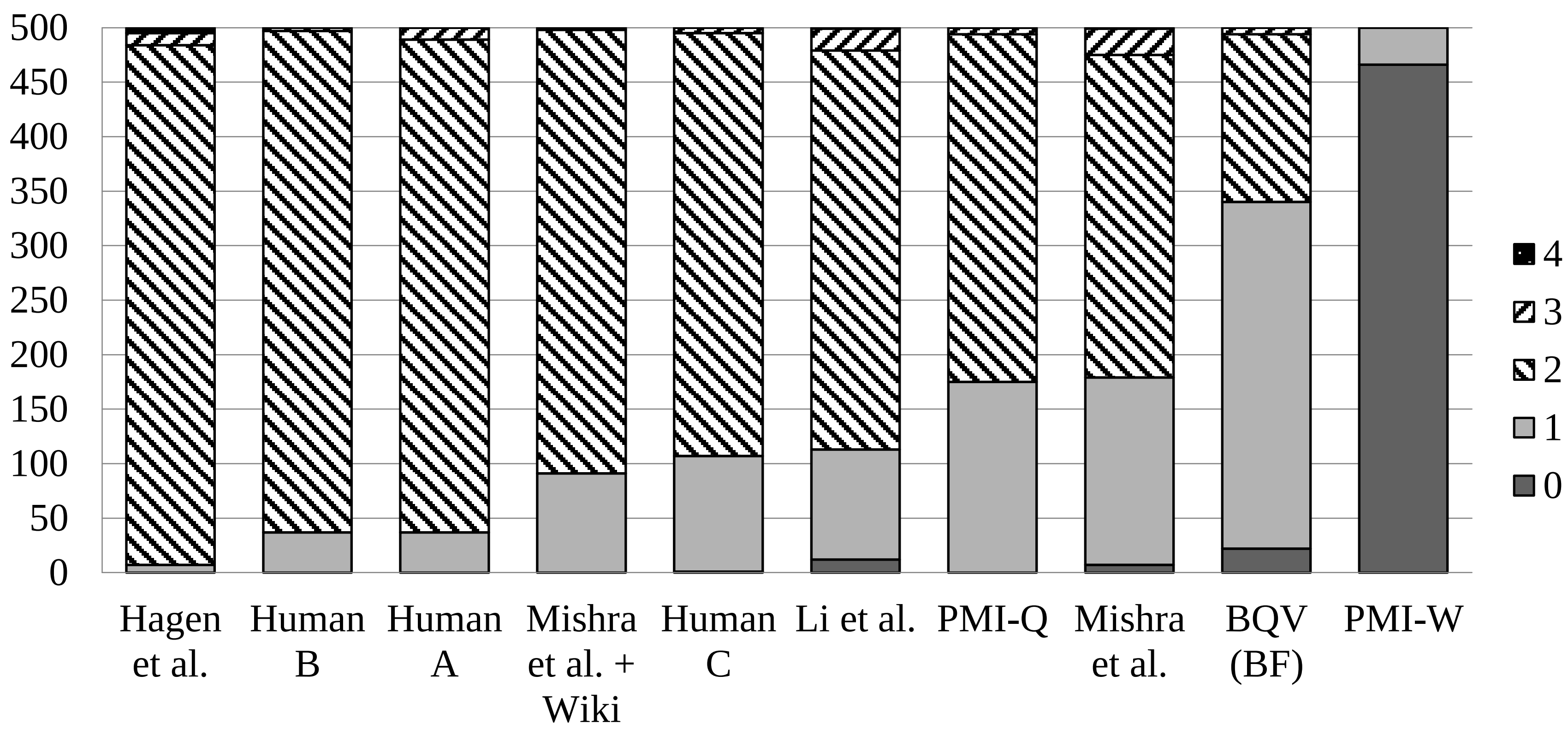
Results

- Kendall-Tau between rankings derived
- IR-performance and Matching Metrics (Humans as reference): 0.75
  - Crucial rank inversions for certain pairs when performances compared (Li et al. and PMI-Q)
- IR-performance and Matching Metrics (BQV (BF) as reference): – 0.85
  - Issues with metrics!

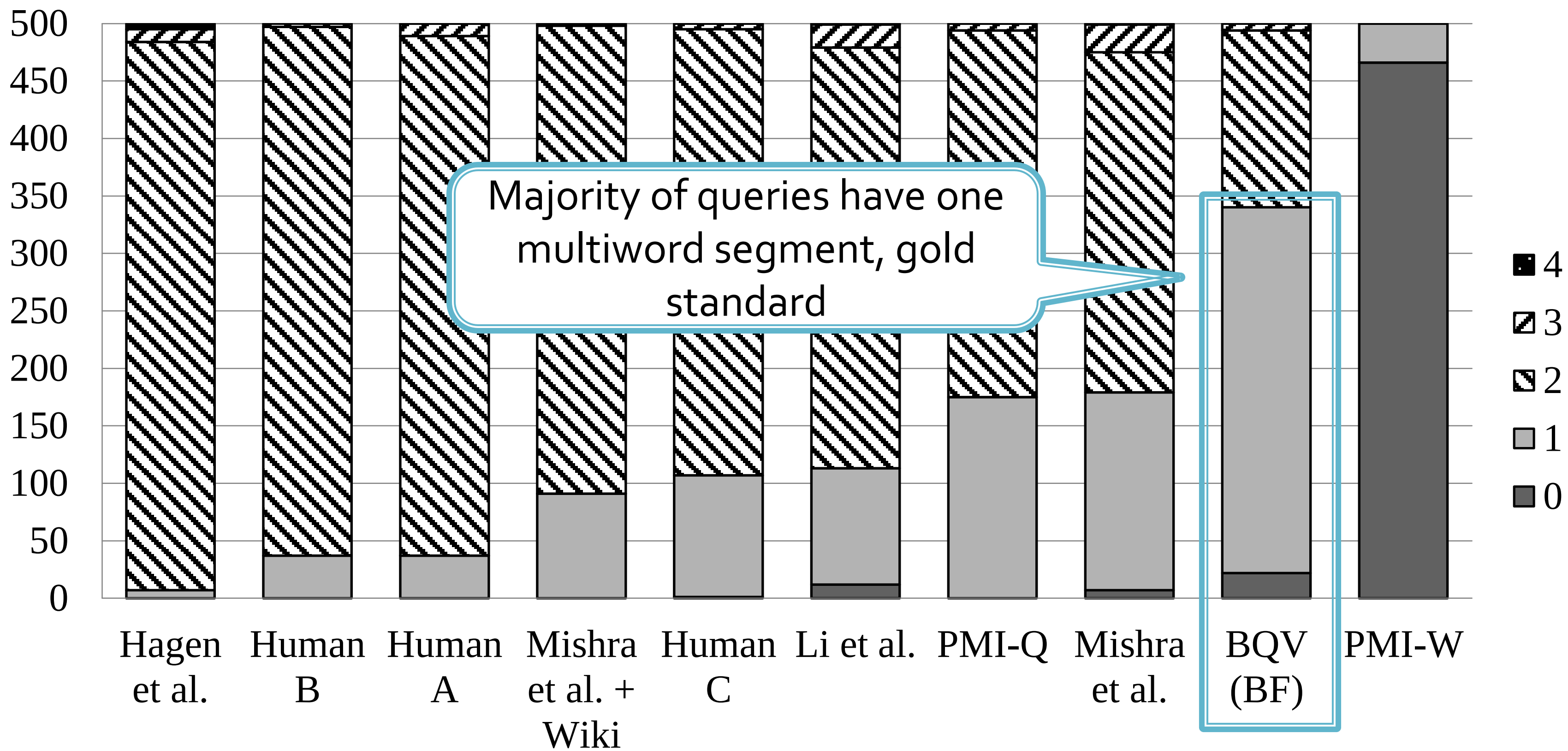
Results

- **Algo. 1:** *history | of | all saints | church | south australia*
- **Algo. 2:** *history of all | saints church south | australia*
- **Human:** *history of | all saints church | south australia*
- **IR-performance:** Algo. 1 > Algo. 2
- **Matching metrics:** Algo. 1 ≈ Algo. 2
  - Sub-, super- and straddle – same penalty for all!

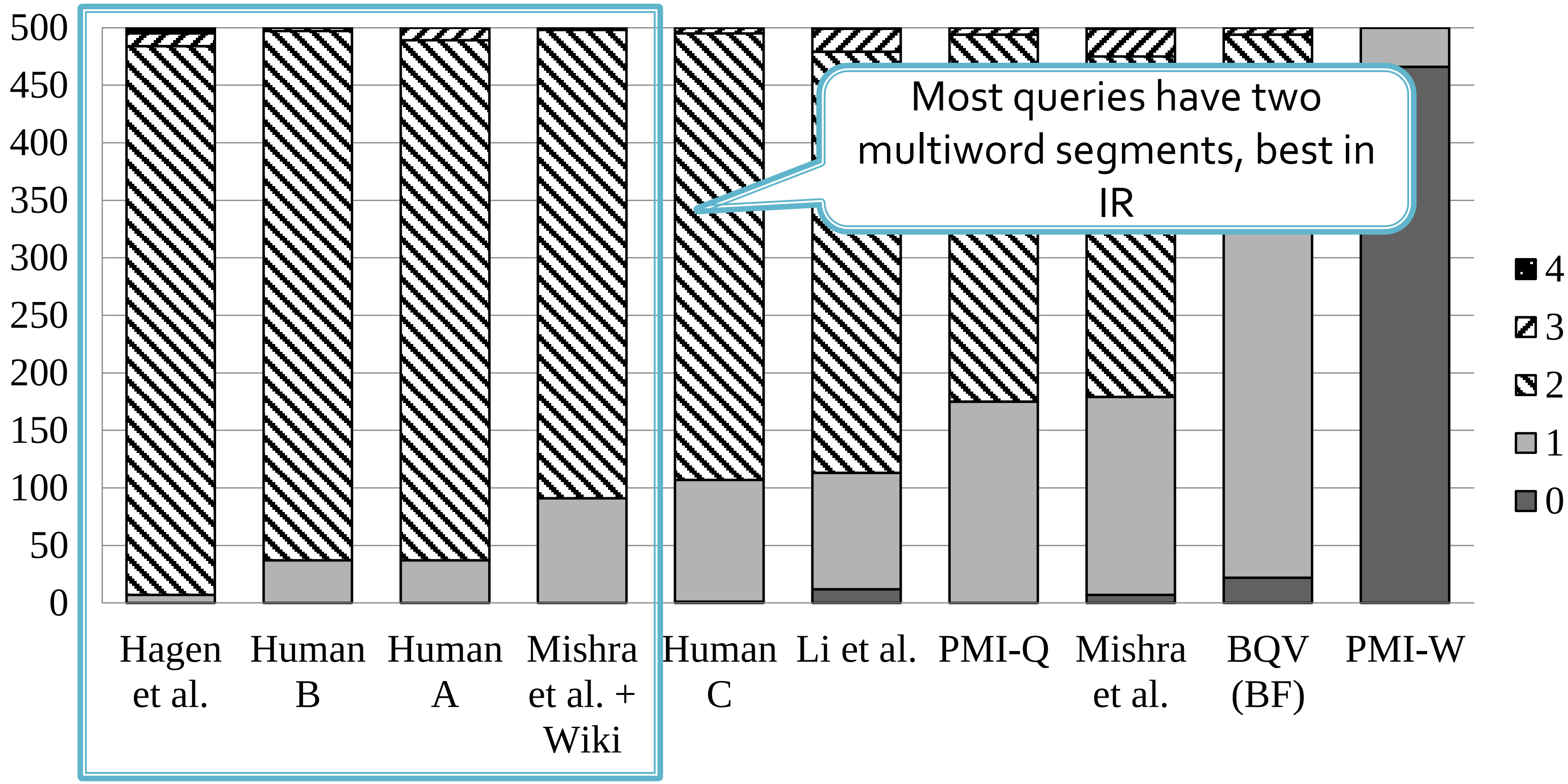
Multiword Segment Analysis



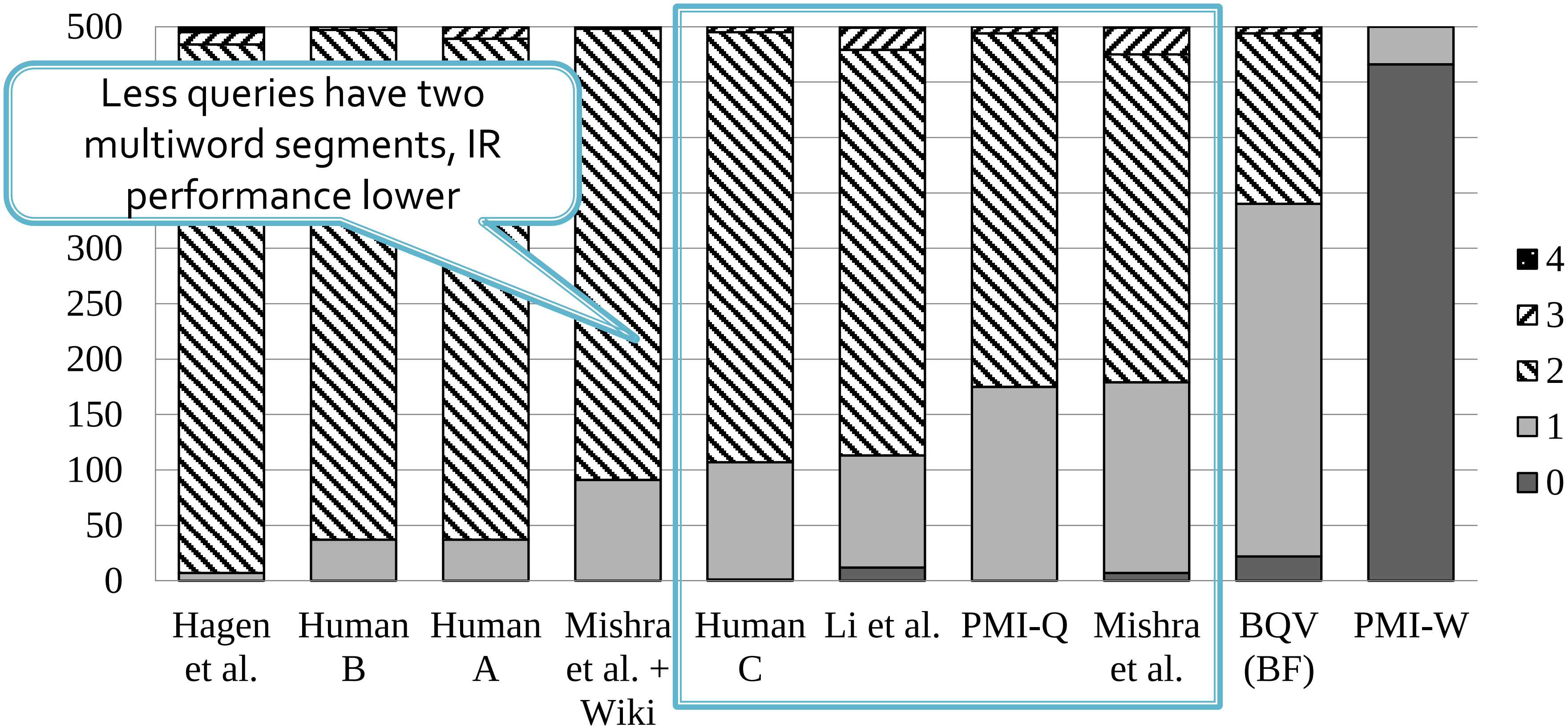
Multiword Segment Analysis



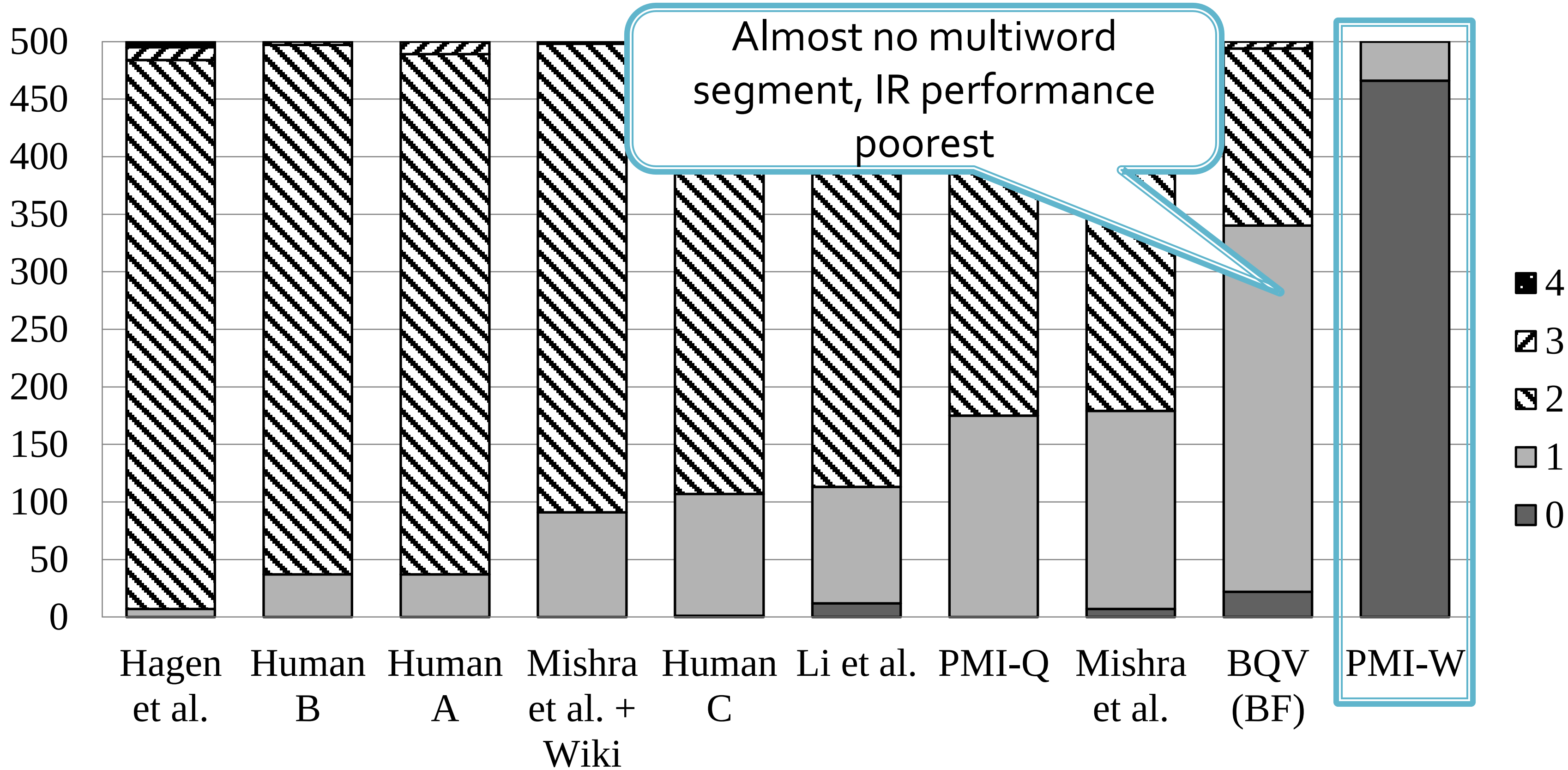
Multiword Segment Analysis



Multiword Segment Analysis



Multiword Segment Analysis



Observations

- Human as well as all algorithmic segmentation schemes consistently outperform unsegmented queries
- Performance of some segmentation algorithms are comparable and sometimes even marginally better than some of the human annotators
- Considerable scope for improving IR performance through better segmentation (all values less than BQV (BF))

Insights

- Segmentation is helpful for IR
- Human segmentations are a good proxy, but not a true gold standard
- Matching metrics are misleading – no differential penalties
- Distribution of multiword segments across queries gives insights about effectiveness of strategy
  - Vital for algorithms to detect multiword segments that are important for IR – output should allow the BQV(BF) to be generated

Final words

- Dataset used for all experiments publicly shared at

http://cse.iitkgp.ac.in/resgrp/cnerg/qa/querysegmentation.html

- **Acknowledgements:**
  - ACM SIGIR Student Travel Support and the Donald B. Crouch Travel Grant
  - Microsoft Research Ph.D. Fellowship
  - Matthias Hagen (Bauhaus Universitat Weimar) for providing us with the segmentation output of his segmentation algorithm (Hagen et al., 2011)
  - Kuansan Wang and Bo-June (Paul) Hsu (Microsoft Research Redmond) for sharing the code for their segmentation algorithm (Li et al., 2011)

Questions?


Thank You!

Backup Slides

- 500 queries resulted in 4,476 quoted versions (approx. 9 per query)
- Fetched 14, 171 unique URLs (approx. 28 per query, 3 per quoted version)
- On an average, adding the 9th strategy to a group of the remaining eight resulted in about one new quoted version for every two queries
- These new versions may or may not introduce new documents to the pool

Backup Slides

- For 71.4% of the queries there is less than 50% overlap between the top ten URLs retrieved for the different quoted versions

Backup Slides

Metric	Unseg.	[11]	[6]	[13]	[13] + Wiki	PMI-W	PMI-Q	*A*	*B*	*C*	*BQV*
nDCG@5	0.688	0.752*	**0.763***	0.745	**0.771***	0.691	**0.766***	**0.770**	0.768	0.759	0.802*
nDCG@10	0.701	0.756*	**0.767***	0.751	**0.771***	0.704	**0.767***	**0.770**	**0.768**	**0.763**	0.813*
MAP@5	0.882	0.930*	**0.942***	0.930*	**0.946***	0.884	0.932*	**0.944**	**0.942**	0.936	0.950*
MAP@10	0.865	0.910*	**0.921***	0.910*	**0.924***	0.867	0.912*	**0.923**	**0.921**	**0.916**	0.935*
MRR@5	0.538	0.632*	**0.649***	0.609	**0.657***	0.543	**0.648***	**0.656**	**0.648**	0.632	0.716*
MRR@10	0.549	0.640*	**0.658***	0.619	**0.665***	0.555	**0.656***	**0.665**	**0.656**	0.640	0.724*

- **IR performance** of state-of-the-art schemes ([11] – Li et al. (SIGIR 2011), [6] – Hagen et al. (WWW 2011), [13] – Mishra et al. (WWW 2011))

- BQV stands for the best quoted version. The highest value in a row (excluding the BQV column) and those with no statistically significant difference with the highest value are marked in boldface. The values for algorithms that perform better than or have no statistically significant difference with the minimum of the human segmentations are marked with *. The paired t-test was performed and the null hypothesis was rejected if the p-value was less than 0.05.

Backup Slides

Metric	Unseg	[13]	[8]	[16]	[16] + Wiki	PMI-W	PMI-Q	*A*	*B*	*C*	*BQV*
Qry-Acc	0.000	0.375	0.602*	0.167	**0.749***	0.000	0.341	0.631	0.686	0.589	0.065
Seg-Prec	0.043	0.524	0.697*	0.350	**0.803***	0.036	0.448	0.691	0.741	0.682	0.140
Seg-Rec	0.076	0.588	0.713*	0.447	**0.785***	0.059	0.487	0.714	0.766	0.723	0.170
Seg-F	0.055	0.554	0.705*	0.392	**0.794***	0.045	0.467	0.702	0.753	0.702	0.153
Seg-Acc	0.404	0.810	0.885	0.748	**0.927***	0.411	0.810	0.892	0.913	0.893	0.654

- The highest values in a row with no statistically significant differences between each other are marked in boldface. The values for algorithms that perform better than or have no statistically significant difference with the minimum of the values for human segmentations are marked with *. The paired t-test was performed and the null hypothesis was rejected if the p-value was less than 0.05.
- **Performance** of state-of-the-art schemes against manual segmentations (Bing test set)
- Crucial inversions of ranks of PMI-Q and [13]

Backup Slides

Table 7: IR-based evaluation using Bing API.

Metric	Unseg. query	All quoted for [11] + Wiki	Oracle for [11] + Wiki
nDCG@10	0.882	0.823	**0.989***
MAP@10	0.366	0.352	**0.410***
MRR@10	0.541	0.515	**0.572***

The highest value in a row is marked **bold**. Statistically significant ($p < 0.05$ for paired t-test) improvement over the unsegmented query is marked with *.