The Effect of Expanding Relevance Judgements with Duplicates

Gaurav Baruah, Adam Roegiest, Mark D. Smucker

University of Waterloo

BACKGROUND: The TREC 2013 Temporal Summarization Track (TST) had systems return relevant sentences between two points in time from a time-ordered document stream.

OBSERVATION: There are many duplicates of judged and unjudged sentences in the query time periods.

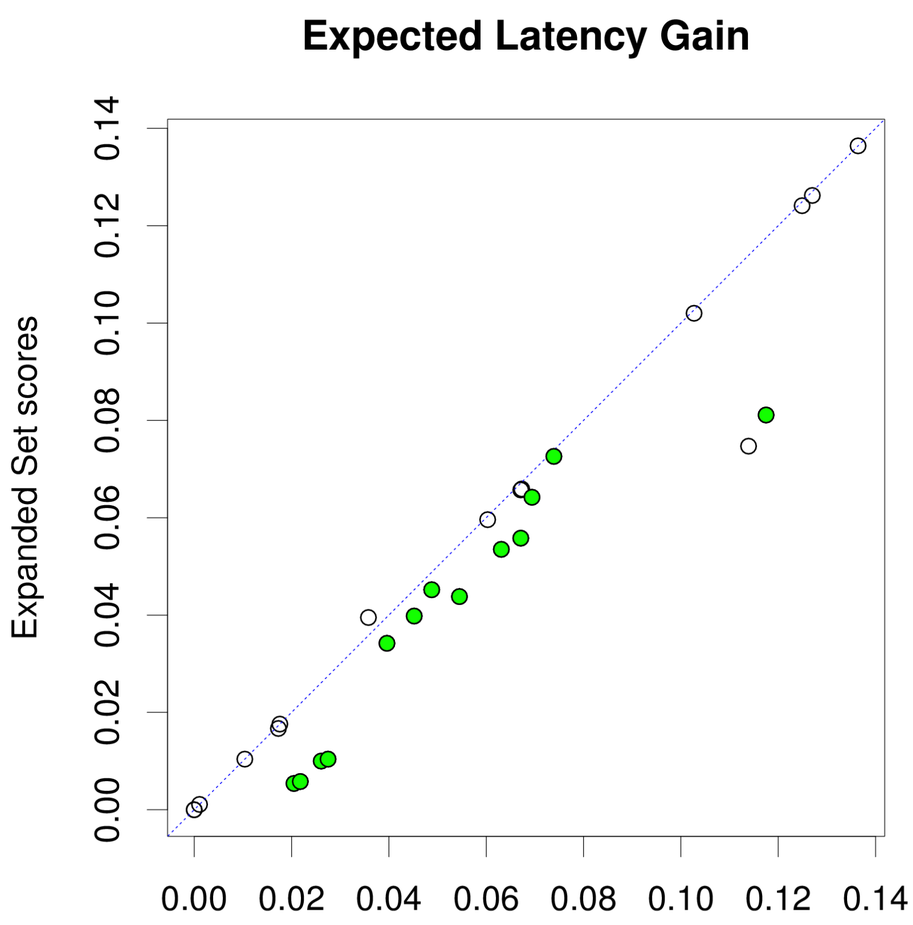
EVALUATION POOL		EXACT DUPLICATES	
#sentences	#relevant	#sentences	#relevant
9,113	2,635	9,025,066	94,621

COUNT	SENTENCE (Text Content)	R/NR
3,376,809	All rights reserved./All rights reserved	NR
2,013,684	Yahoo!	NR
673,876	New User ?	NR
...		
294,662	This material may not be published, broadcast, re-written or redistributed.	NR
...		
5,403	National Hurricane Center in Miami said Isaac became a Category 1 hurricane Tuesday with winds of 75 mph.	**R**
...		

QUESTION: Does the inclusion of duplicate sentences affect the evaluation of the TST?

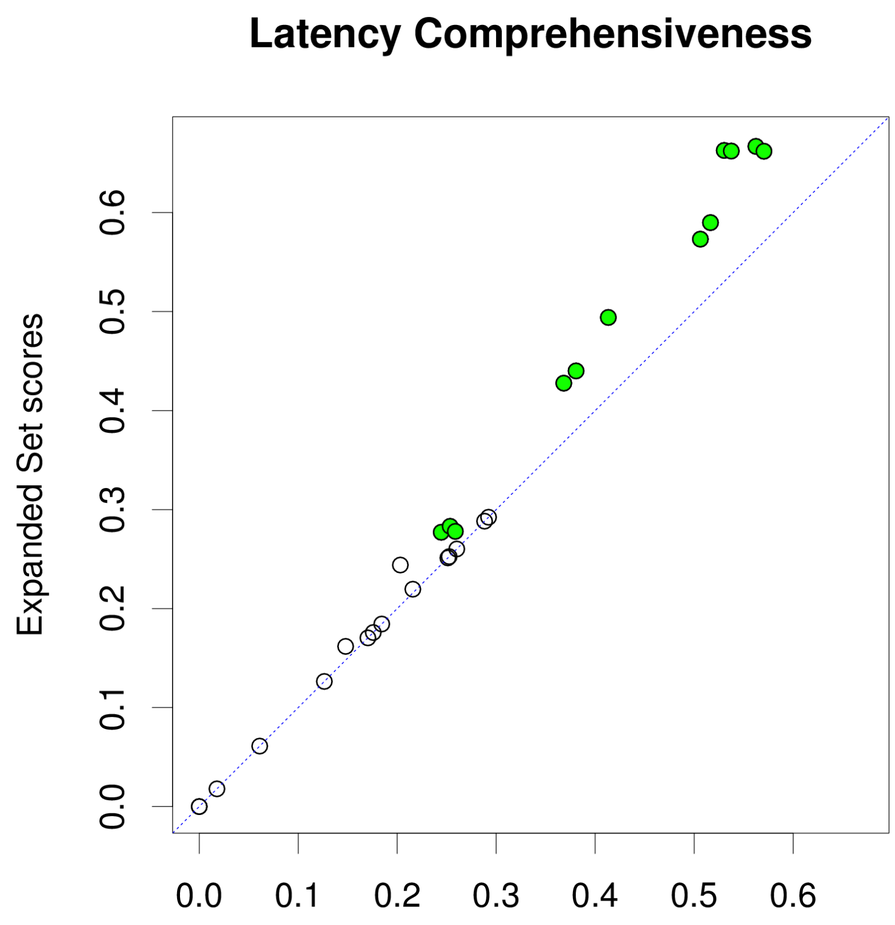
EXPERIMENT: Expand the original judged set with duplicates and compare system performance over the TST metrics for the 28 runs submitted to the track.

RESULTS:



Rank Correlation Kendall's τ = 0.899.

Statistically significant score change for 13 runs.



Rank Correlation Kendall's τ = 0.942.

Statistically significant score change for 12 runs.

CONCLUSION: The inclusion of duplicate sentences yields a high correlation with the original rank order, but it changes the scores of many systems. We recommend adding duplicate sentences to the judged set, which could help enhance test collection reusability.

ACKNOWLEDGEMENT: Support provided in part by a SIGIR 2014 Student Travel Grant.