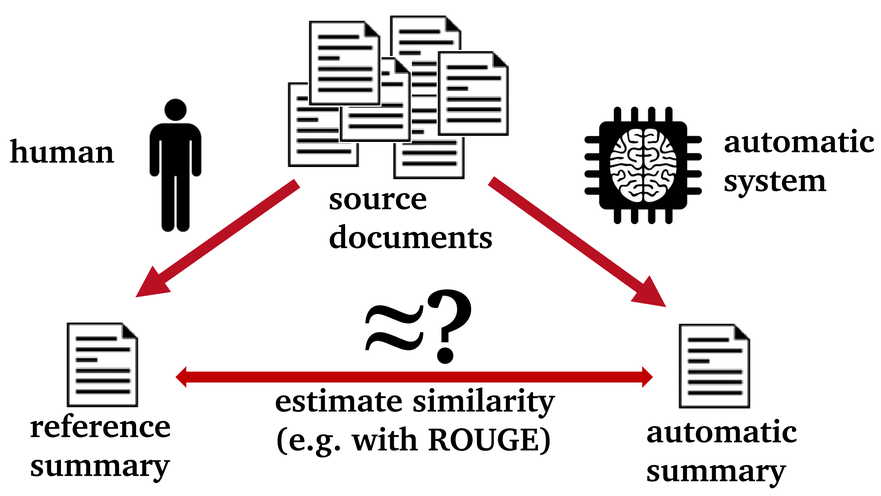
Estimating Summary Quality with Pairwise Preferences

Markus Zopf | Research Training Group AIPHES, TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Traditional Automatic Evaluation



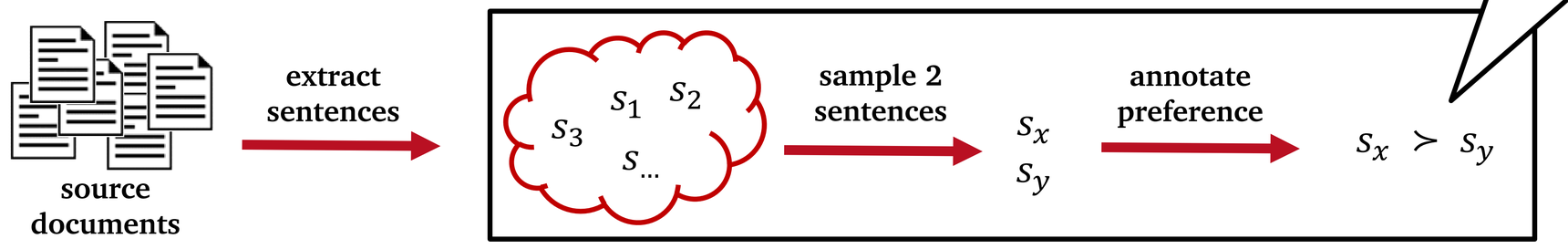
Problems:

1. creating reference summaries is time-consuming and complex → expensive
2. computing text similarities is a complex, unsolved problem → not reliable

Summary

- traditional evaluation methods use *references summaries* to estimate the quality of automatically generated summaries
- creating references summaries is *complex and expensive*
- we propose to use *simple pairwise preference labels* of individual sentences for evaluation
- our evaluation is *cheaper and more accurate* than the SOTA

Estimating Summary Quality with Pairwise Preferences



generated preferences are interpreted as games in which better sentences are more likely to win a game

$s_4 \succ s_6$
$s_2 \succ s_4$
$s_8 \succ s_5$
...

compute Bradley-Terry scores

$$p(s_x \succ s_y) = \frac{u(s_x)}{u(s_x)+u(s_y)}$$

$$v(\mathbf{S}) = \sum_{i}^{|\mathbf{S}|} w_i \cdot u(s_i)$$

S=summary to evaluate
s_i=i-th sentences in **S**

Automatically Generating Preferences

reference summaries are already available for standard datasets → reuse reference summaries to automatically create preference annotations for free

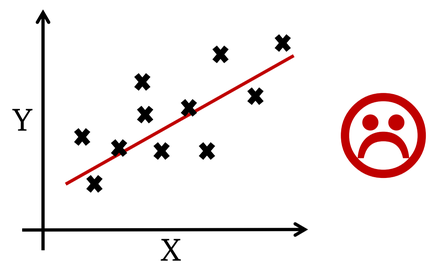
reference summary

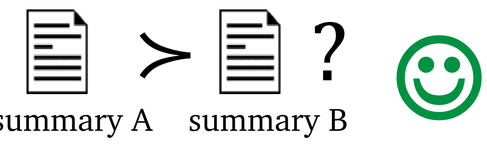
$$\hat{u}(s_i) = \max_{s \in \text{ref sum}} \text{sim}(s_i, s)$$

$s_x \succ s_y$

automatically generated preference label

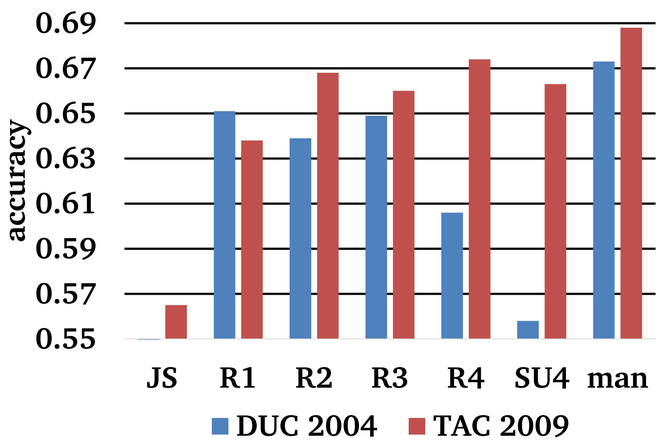
Evaluating Evaluation Methods





Prior work: Pearson correlation	Our work: Pairwise accuracy
interpretation difficult	clear interpretation
requires linear correlation	only necessary requirement
normal distr., interval scaling assumed	no additional assumptions
sensible to outliers	robust

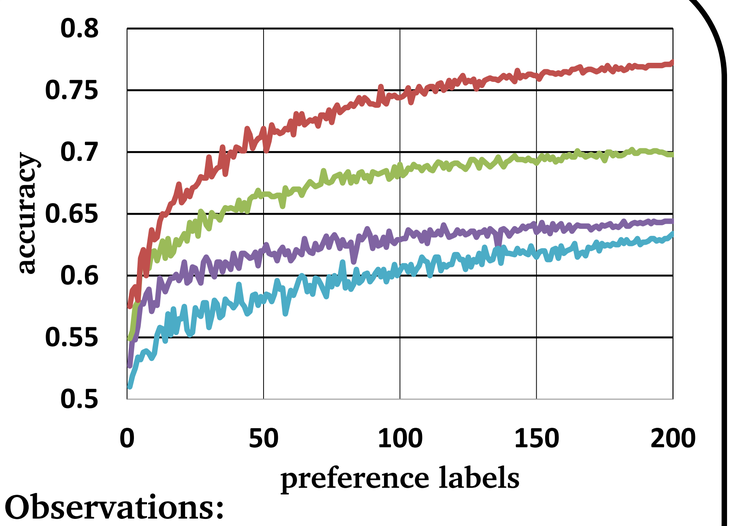
Using 200 Human Preferences



Observations:

- annotation effort: ~54 minutes per topic
- preference-based evaluation with human annotations outperforms JS and ROUGE

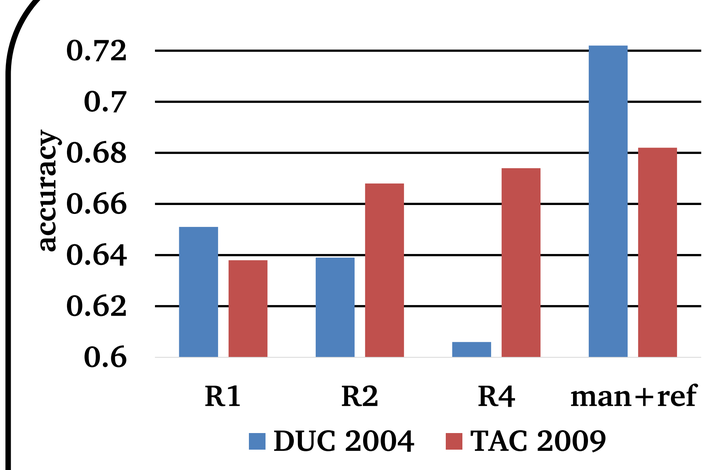
Training Convergence



Observations:

- different topics → different convergence
- not yet converged after 200 labels

Human+Automatic Preferences



Observation:

- using additional automatically generated labels can further improve accuracy

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1

www.aiphes.tu-darmstadt.de