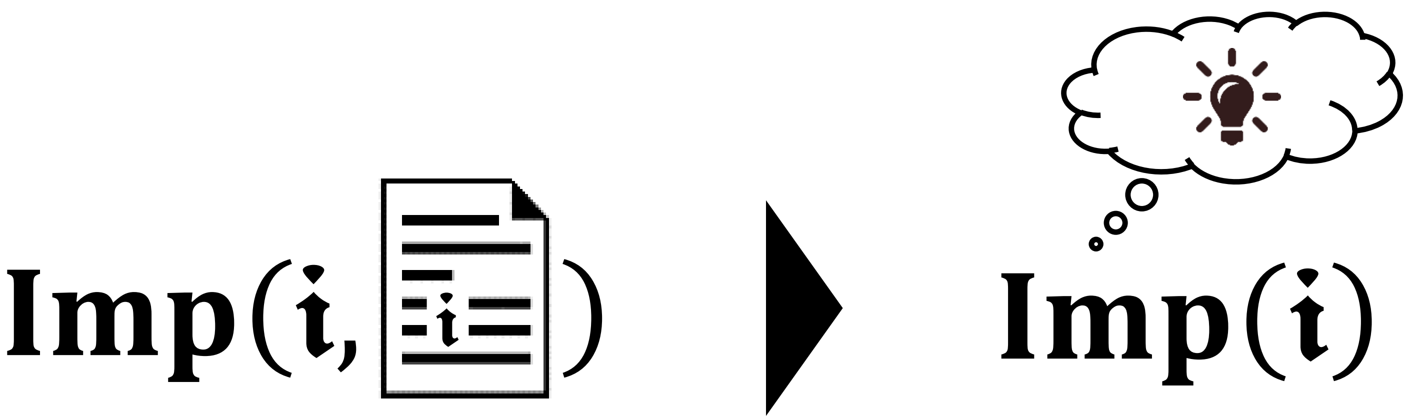
Towards Context-free Information Importance Estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

Markus Zopf
zopf@aiphes.tu-darmstadt.de
https://www.aiphes.tu-darmstadt.de



Automatic Information Importance Estimation

Need for Automatic Information Importance Estimation (IIE)

Information Overload

Importance of Information

Towards Context-free Information Importance Estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

Evaluation

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

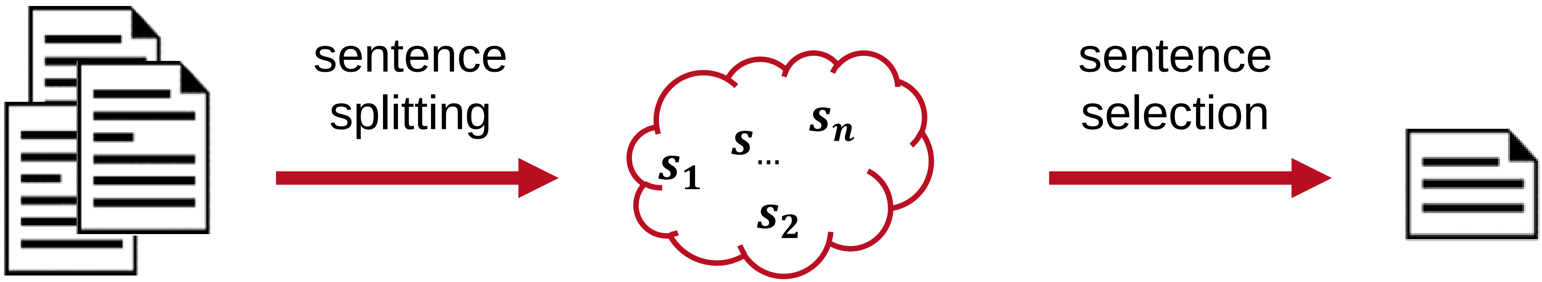
Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

Evaluation

(Extractive) Summarization



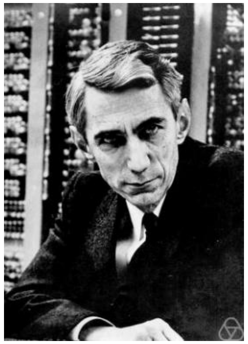
1. make it shorter: $|\text{output}| < \sum_{i=1}^{n} |\text{input}_i|$
2. don't add content: $\mathfrak{c}(\text{output}) \subseteq \bigcup_{i=1}^{n} \mathfrak{c}(\text{input}_i)$, $\mathfrak{c} : \text{Text} \rightarrow \text{Content}$
3. maximize a utility: $\text{output}^* = \underset{\text{output} \in \text{Text}}{\operatorname{argmax}} \; u\big(\mathfrak{c}(\text{output})\big)$

Definitions of Summarization

source	major parts / important information
Wikipedia	✓
Oxford Dictionaries	✓
[Cleveland et al., 2013]	✓
[Spärck Jones, 1999]	✓
[Radev, et al., 2002]	✓
[Hovy, 2005]	
[Torres-Moreno, 2014]	✓
[Saggion et al., 2002]	
[Mani, 2001]	✓

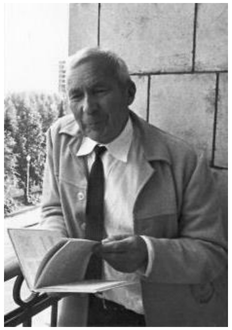
But: Which information is important?

Prior Methods for Information Value Estimation



Shannon Information

$\mathrm{Inf}_{\mathrm{Sh}}(w_i) = -\log(\Pr(w_i))$



Kolmogorov Complexity

$\mathrm{Inf}_{\mathrm{Ko}}(s)$ = “length of shortest algorithm to compute *s*”

„I give a talk today“ vs. „Ffps gkw bci flvmnd eit“

Definition: Information = Data + Meaning

Luciano Floridi
"Philosophy of Information"

"The element $\mathfrak{i}$ is an instance of information, understood as semantic concept, if and only if

1. *$\mathfrak{i}$ consists of n data with $n \geq 1$,*
2. *the data are well-formed, and*
3. *the well-formed data are* ***meaningful****"*

meaningful data ≈ data that can be interpreted/understood

„I give a talk today" vs. „Ffps gkw bci flvmnd eit"

[Bar-Hillel & Carnap,1953; Floridi, 2010]

Amount of Semantic Information

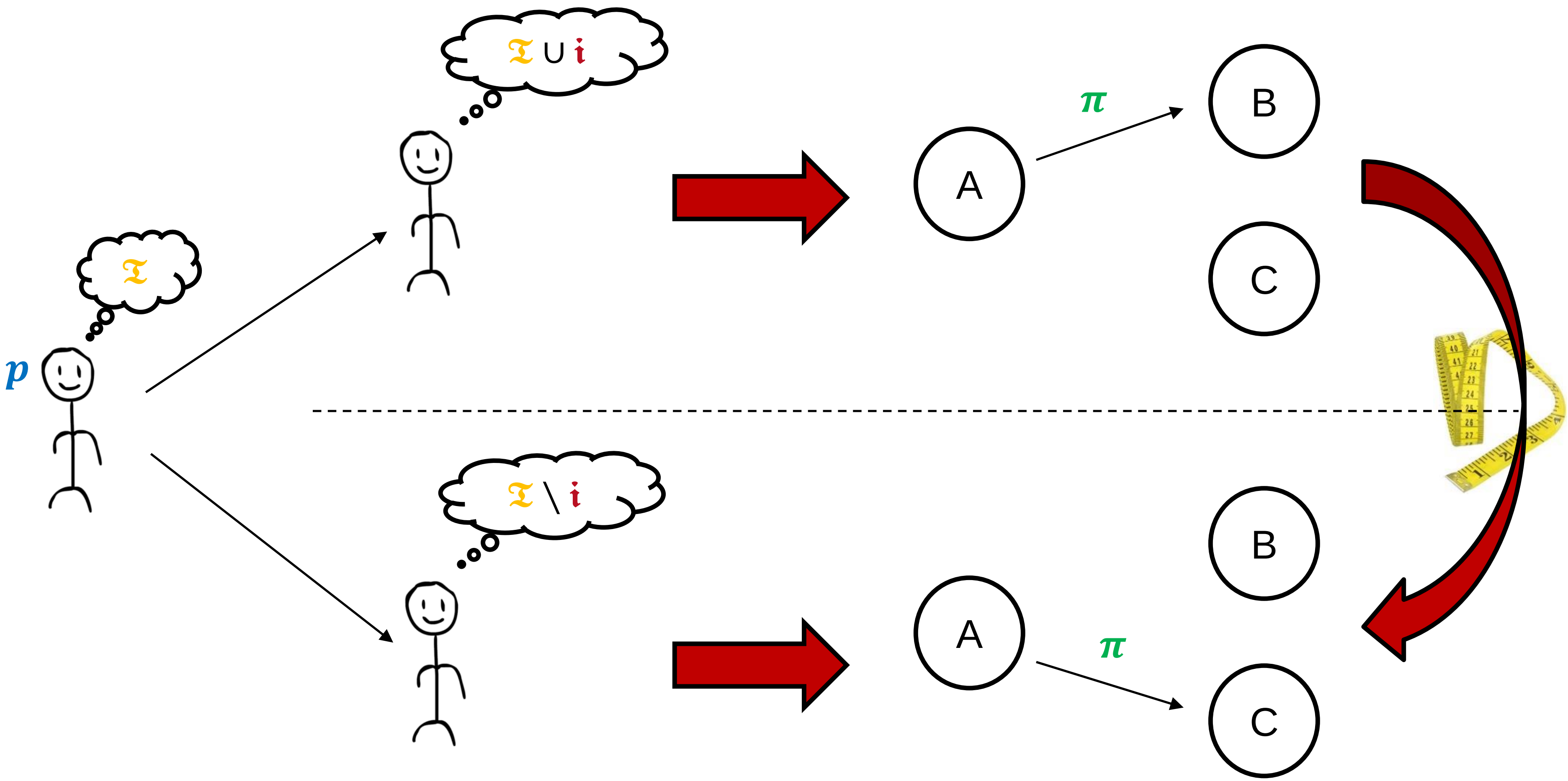
Semantic Information

$$\mathrm{Inf}_{Sem}(\mathfrak{i}) = |S| - |S_i|$$

- system $\mathcal{S}$
- S = prior knowledge about $\mathcal{S}$
- S_i = knowledge about $\mathcal{S}$ after consuming information $\mathfrak{i}$

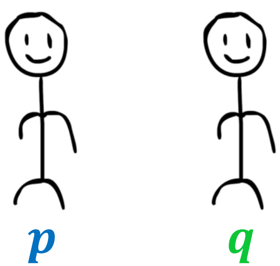
[Floridi, 2010]

A Definition for Information Importance

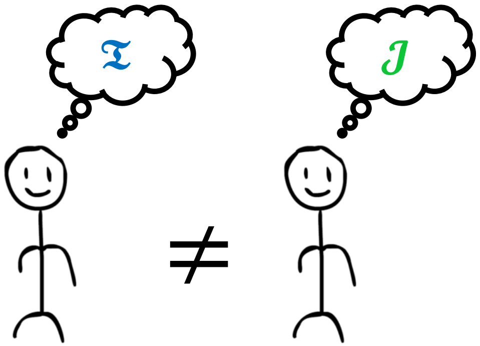


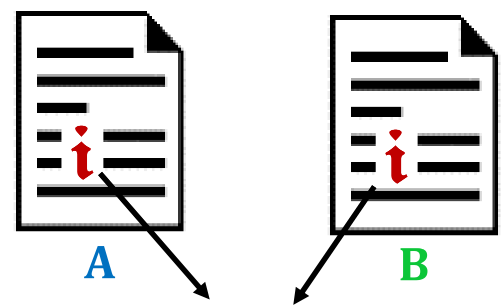
$$\mathrm{Imp}_p(\mathfrak{i}) \;=\; d(\Pr(s|\pi_{\mathfrak{T}\cup\mathfrak{i}}), \Pr(s|\pi_{\mathfrak{T}\setminus\mathfrak{i}}))$$

Information Importance is…



$\mathrm{Imp}_p(\mathfrak{i}) \neq \mathrm{Imp}_q(\mathfrak{i})$





$\mathrm{Imp}_p(\mathfrak{i})$

personalized: $\mathrm{Imp}_p(\mathfrak{i})$

depends on prior knowledge

invariant of the context it appears in: $\mathrm{Imp}_p(\mathfrak{i})$

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

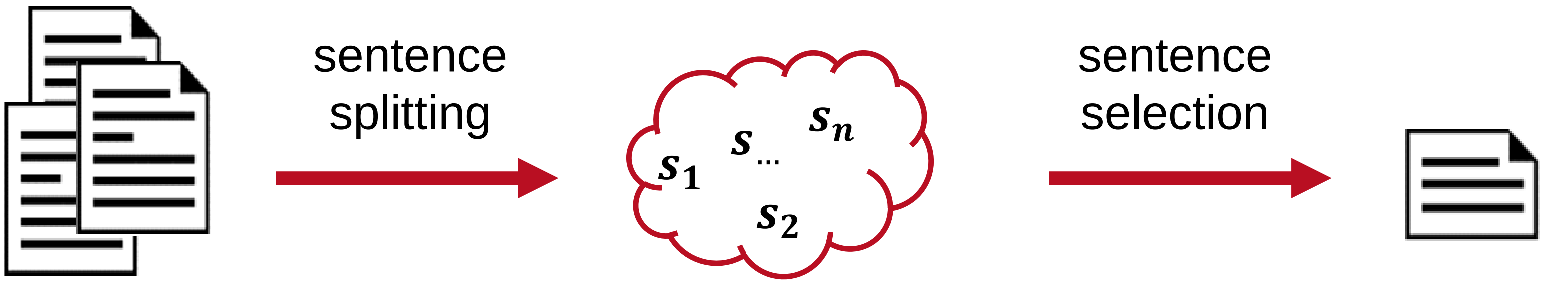
Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

Evaluation

(Extractive) Summarization



1. make it shorter: $|\text{output}| < \sum_{i=1}^{n} |\text{input}_i|$
2. don‘t add content: $\mathfrak{c}(\text{output}) \subseteq \bigcup_{i=1}^{n} \mathfrak{c}(\text{input}_i)$, $\mathfrak{c} : \text{Text} \rightarrow \text{Content}$
3. maximize a utility: $\text{output}^* = \underset{\text{output} \in \text{Text}}{\text{argmax}} \; u\big(\mathfrak{c}(\text{output})\big)$

Disadvantages

- redundancy avoidance + writing quality → additional tasks to be solved jointly
- arbitrary length restrictions → additional parameters/complexity
- only subset selection → coarse-grained (binary) importance estimation

Focus on Importance Estimation

$$\mathrm{i}_j \succ \mathrm{i}_k$$

preference prediction
for information nuggets

$$\begin{aligned}&1.\ \mathrm{i}_j\\&2.\ \mathrm{i}_k\\&3.\ \mathrm{i}_l\end{aligned}$$

information
ranking

$$u(\mathrm{i}_j)$$

utility
prediction

[Zopf et. al, SNAMS 2018]

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

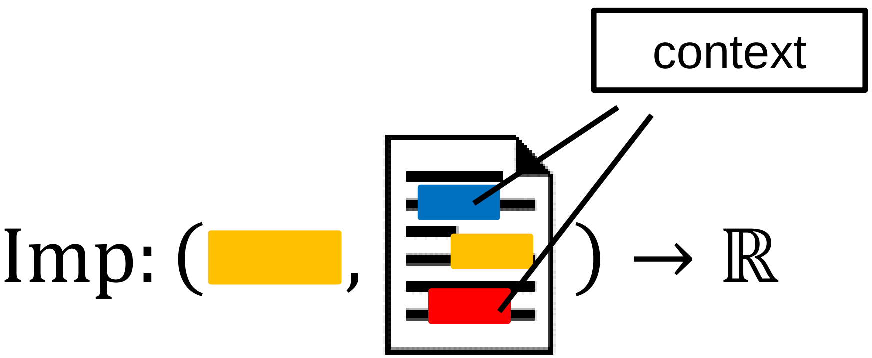
Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

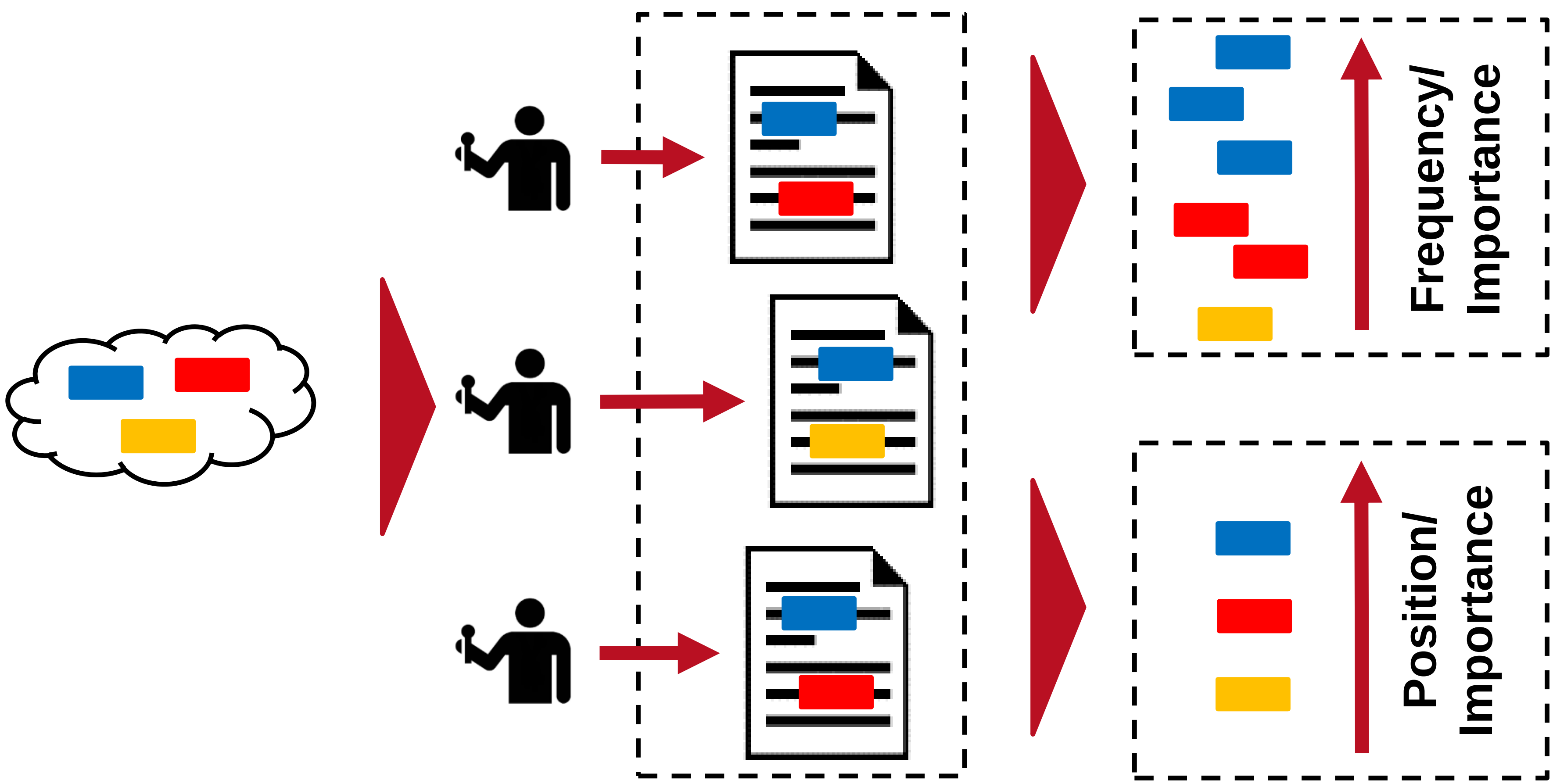
Evaluation

Contextual Importance Estimation

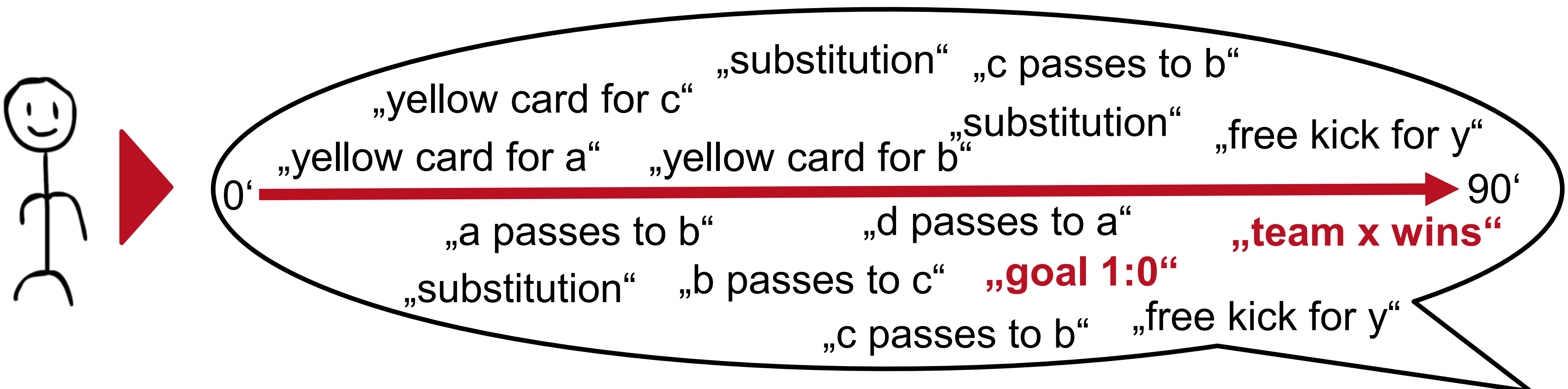


Contextual Information
Importance Estimation

Newswire Articles

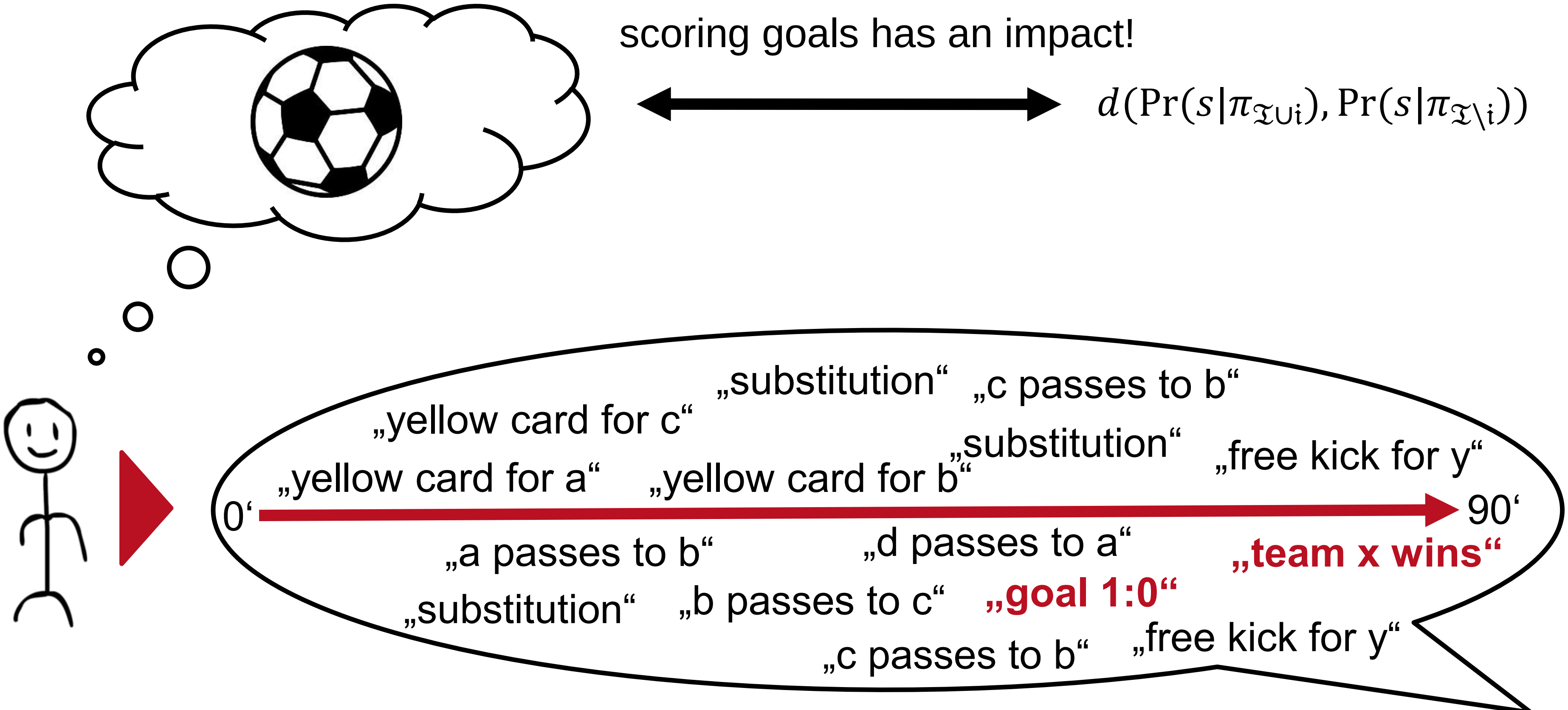


Humans do not need document-derived features to estimate importance



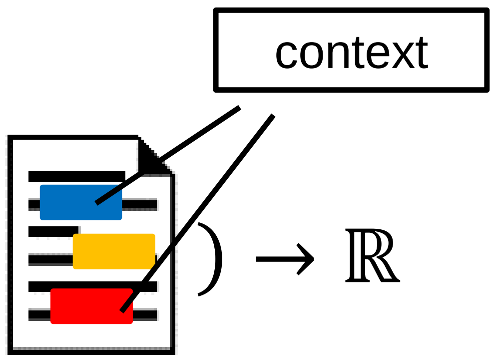
A transcript of a soccer match

Humans use prior domain knowledge for information importance estimation



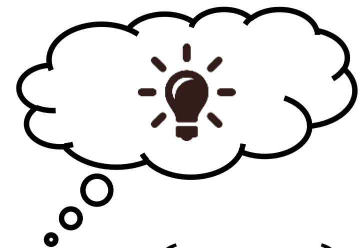
A transcript of a soccer match

Contextual Importance Estimation



$\mathrm{Imp}: (\quad , \quad) \to \mathbb{R}$

Contextual Information Importance Estimation



$\mathrm{Imp}: (\quad) \to \mathbb{R}$

Context-free Information Importance Estimation

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

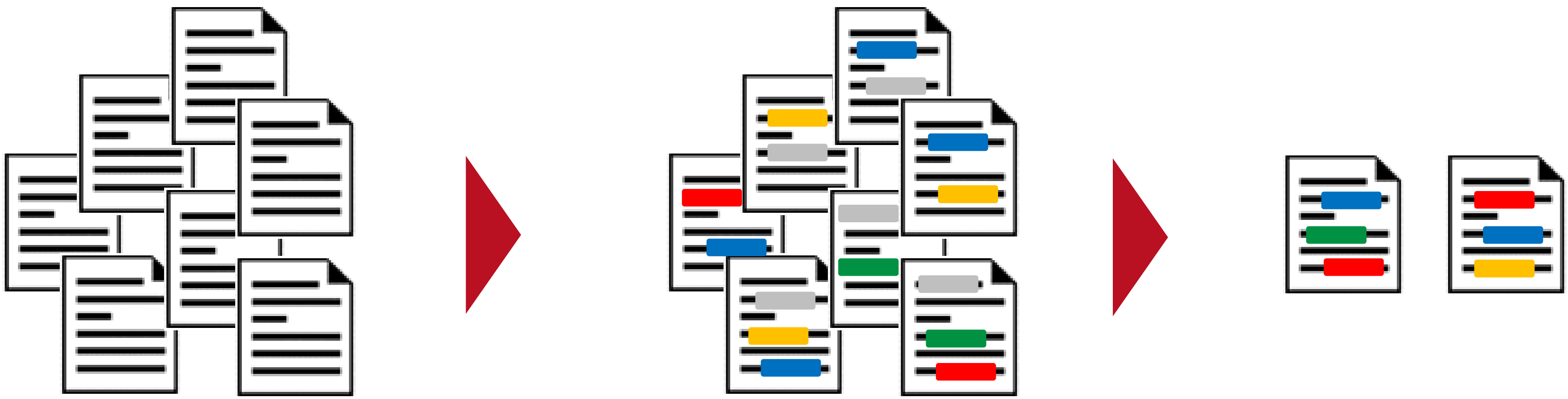
Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

Evaluation

Creating multi-document summarization datasets is complex, tedious, and expensive



search & read source documents

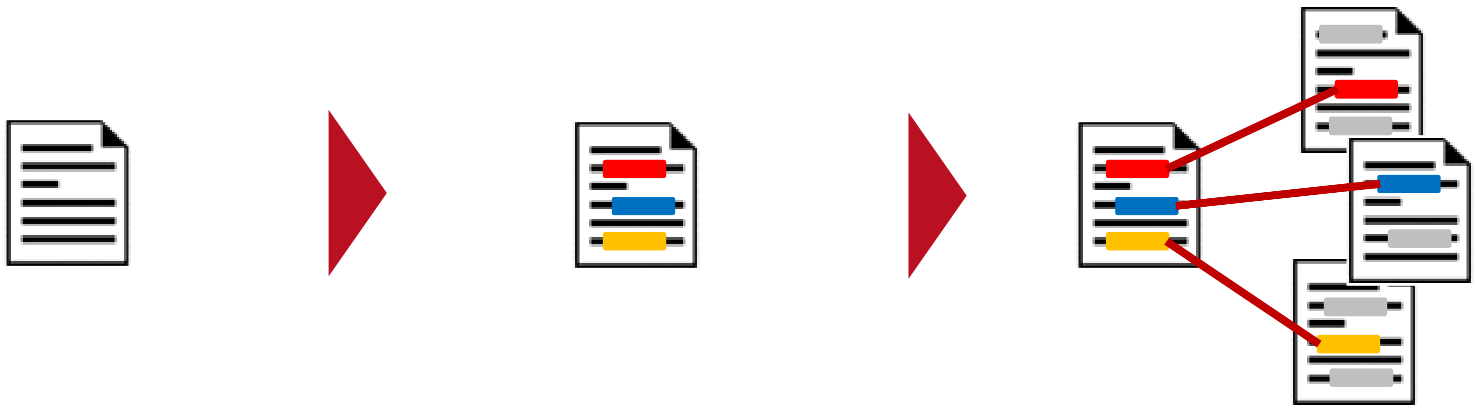
identify important information

write proper summaries

1. finding/reading source documents is **time consuming**
2. identifying important information requires **domain knowledge**
3. professional **writers** required to generate high quality

→ **small multi-document summarization datasets**

A new Corpus Construction Approach



find texts which are summary

annotate all information nuggets

find source for each information nugget

[Zopf et al., COLING 2016]

Wikipedia Provides Summaries

Artikel Diskussion Lesen Quelltext bearbeiten Versionsgeschichte Wikipedia durchsuchen

Bundestagswahlrecht

Das **Bundestagswahlrecht** regelt die Wahl der Mitglieder des Deutschen Bundestages. Nach den in Art. 38 Abs. 1 Satz 1 Grundgesetz (GG) festgelegten Wahlrechtsgrundsätzen ist die Wahl *allgemein, unmittelbar, frei, gleich und geheim*. Das konkrete Wahlsystem wird hingegen durch ein einfaches Gesetz, das Bundeswahlgesetz, bestimmt. Viele Bestimmungen des Bundeswahlgesetzes werden ihrerseits in der Bundeswahlordnung konkretisiert.

Typisch für das deutsche Bundestagswahlrecht ist die Verbindung von Wahlkreiswahl und Listenwahl. Ein Wähler hat zwei Stimmen, eine für einen Direktkandidaten im Wahlkreis und eine für die Landesliste einer Partei. Die Zweitstimme ist entscheidend für den Anteil einer Partei an den Bundestagsmandaten. Gewonnene Wahlkreismandate werden damit verrechnet.

Inhaltsverzeichnis

1 Verfassungsrechtliche Grundlagen
2 Wahlrecht
2.1 Aktives Wahlrecht
2.2 Passives Wahlrecht
3 Wahlorgane

featured articles are

1. well-written
2. comprehensive
3. well-researched
4. neutral
5. stable

lead section

[Zopf et al., COLING 2016]

*h*MDS

- **larger** than previous corpora
- **varying** number of source documents

corpus	# of topics	# of sources / topic ± sd
DUC 2004	50	10 ± 0.0
TAC 2008	48	10 ± 0.0
***h*MDS**	**91**	**13.9 ± 3.1**

- source documents from **many text genres:** articles, forum posts, microblogs, encyclopedic long and short, social media, scientific papers, dialogues
- contains topics in **English and German**
- `github.com/AIPHES/hMDS`

[Zopf et al., COLING 2016]

auto-*h*MDS

- **much larger**

corpus	# of topics
DUC 2004	50
TAC 2008	48
*h*MDS	91
auto-*h*MDS	**7,316**

- **only $0.07** per topic (for search engine)
- contains topics in **English and German**
- `github.com/AIPHES/auto-hMDS`

[Zopf, LREC 2018]

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

Evaluation

approximating unknown function $f\colon \mathcal{X} \rightarrow \mathcal{Y}$

incidental supervision

learn $g \approx f$ without supervision

[Zopf et al., CoNLL 2016]

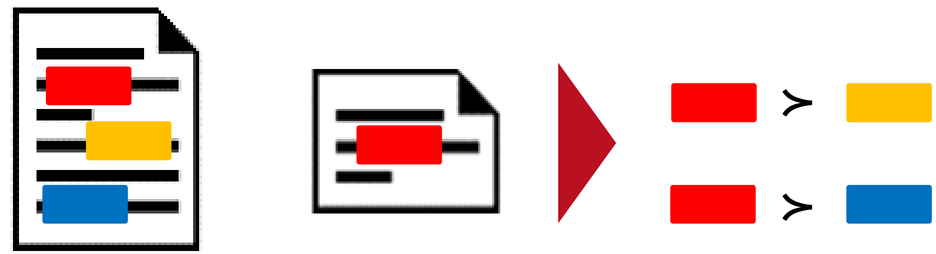
supervised learning (regression)

learn f from $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$

[Zopf et al., NAACL 2018]

Incidental Supervision for Summarization

1. collect preferences:



document D_i summary R_i

2. fit a Bradley-Terry model:

$$\Pr(e_i \succ e_j) = \frac{u(e_i)}{u(e_i) + u(e_j)}$$

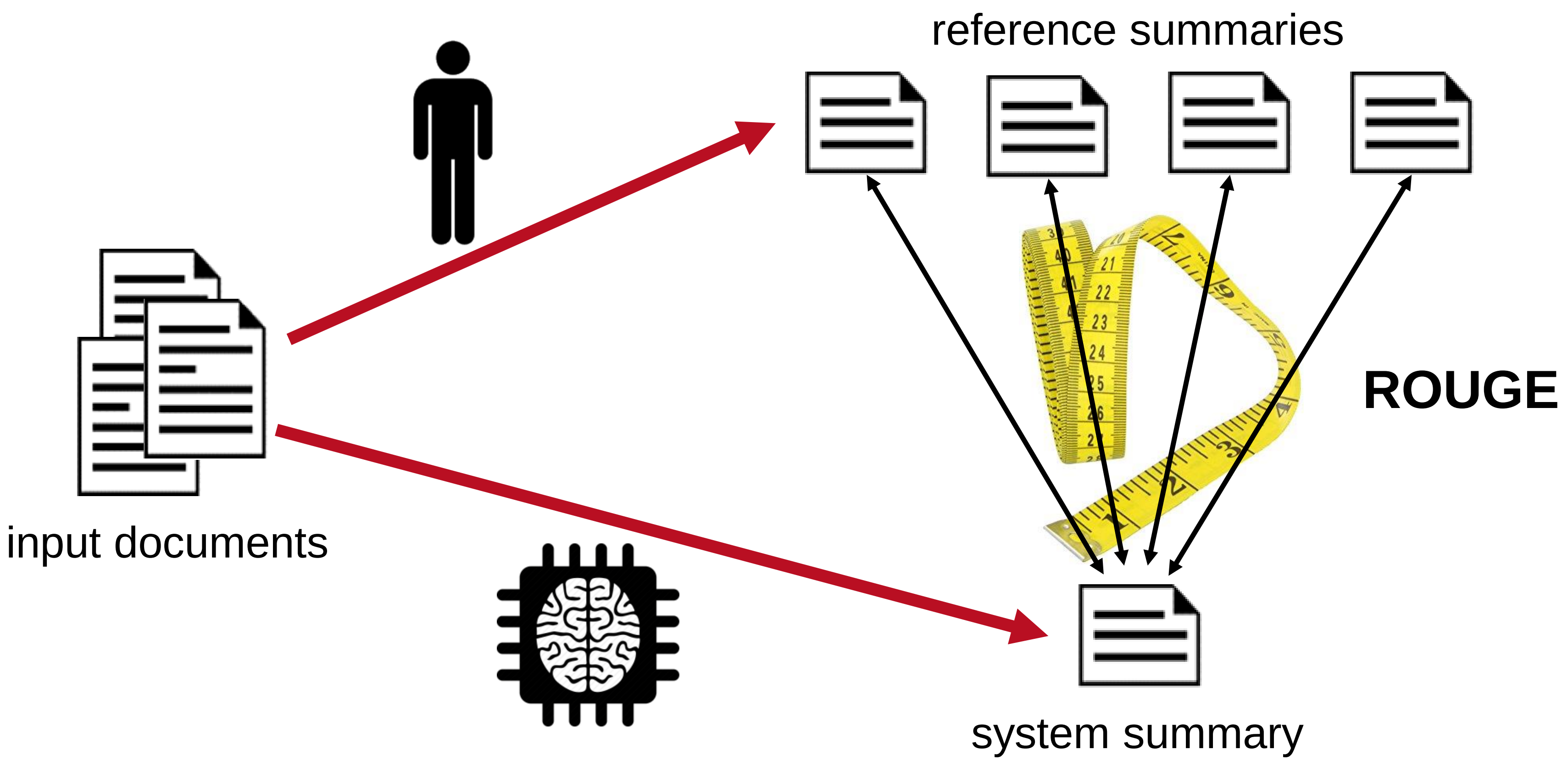
3. estimate sentence utilities:

$$v(s) = \frac{1}{|s|} \sum_{e \in s} u(e)$$

context-free

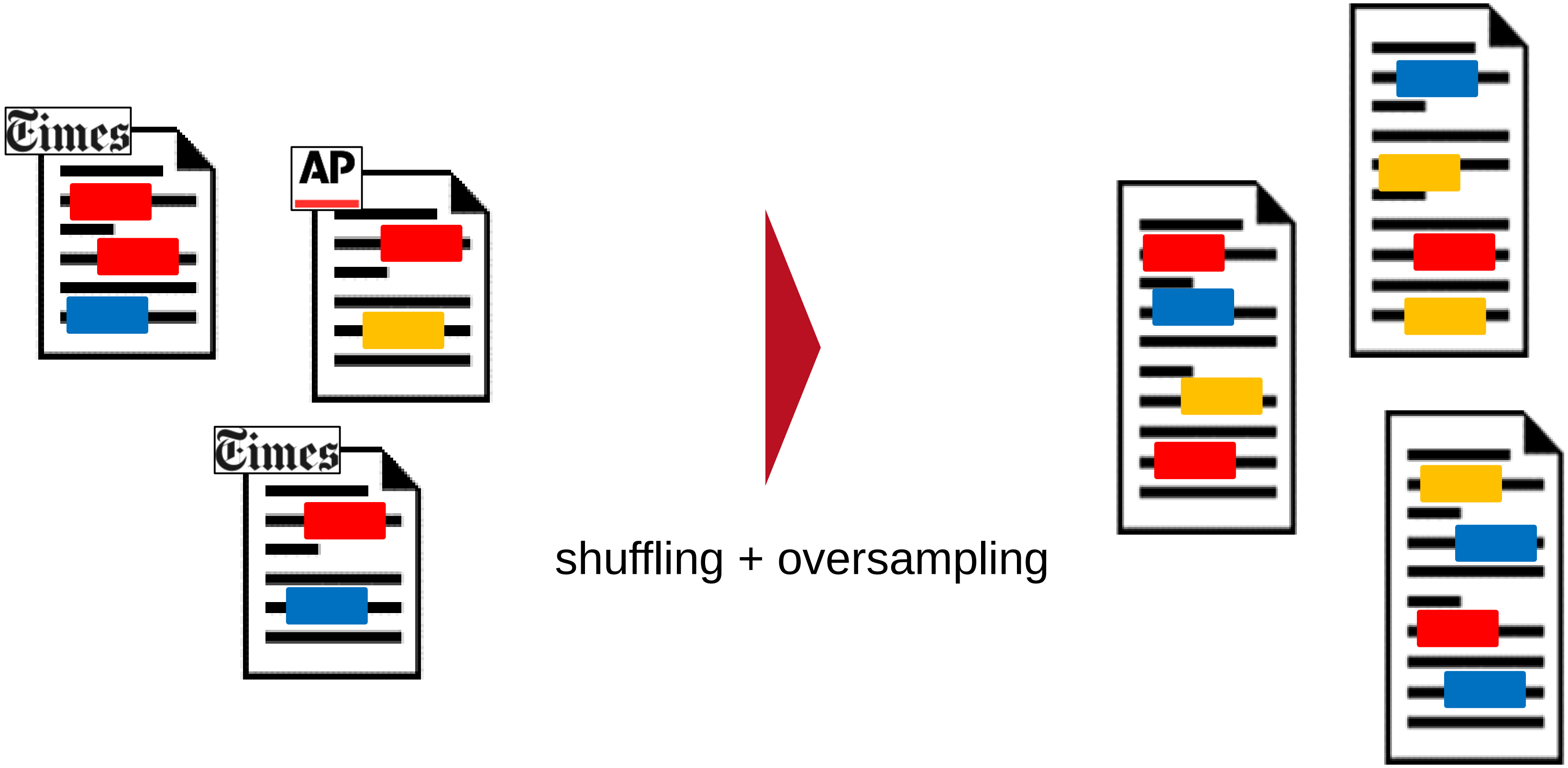
[Zopf et al., CoNLL 2016]

ROUGE



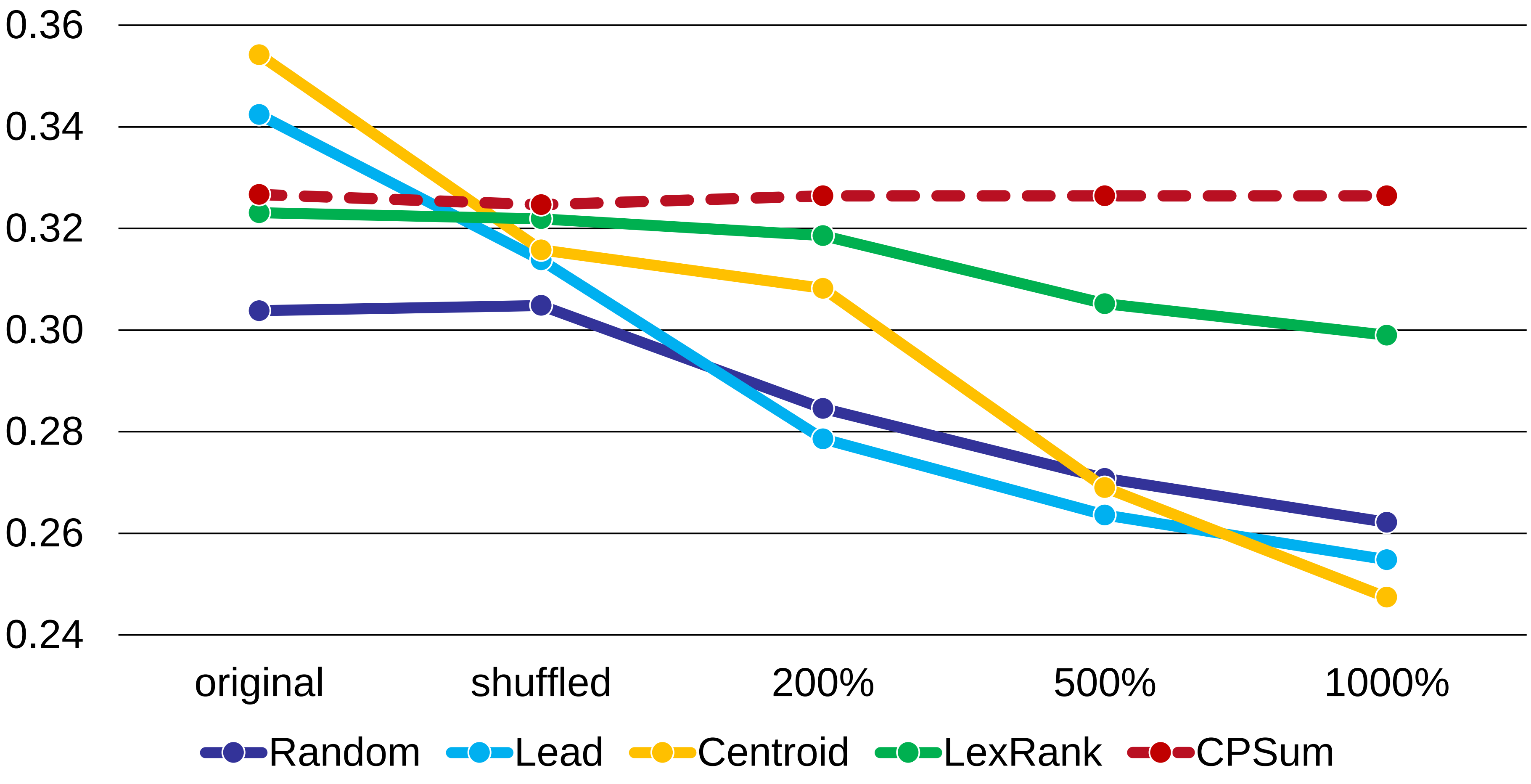
[Zopf et al., CoNLL 2016]

Shuffling and Oversampling



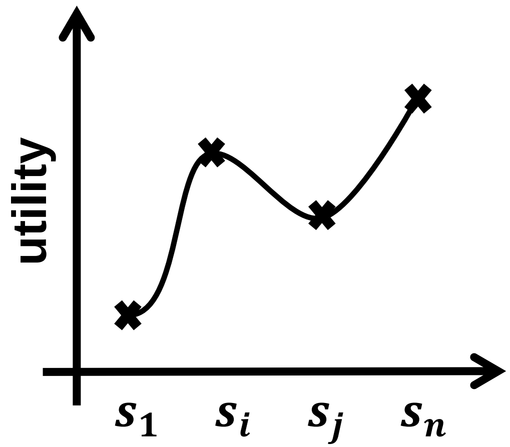
[Zopf et al., CoNLL 2016]

ROUGE-1 scores for different versions of DUC 2004



[Zopf et al., CoNLL 2016]

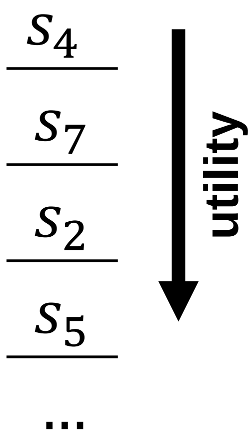
Sentence regression for text summarization



$$f : \varphi(s) \rightarrow \mathbb{R}$$

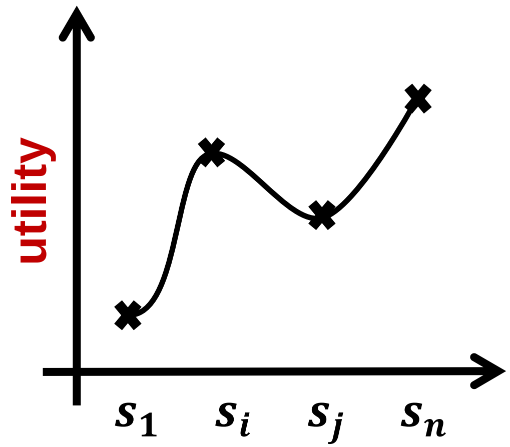
sentence regression

+



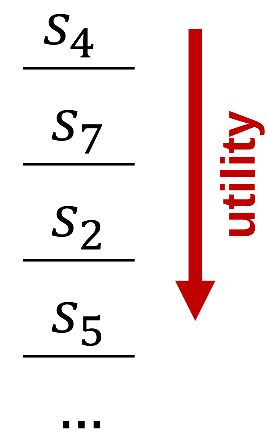
greedy sentence selection

Sentence regression for text summarization



$f : \varphi(s) \rightarrow \mathbb{R}$

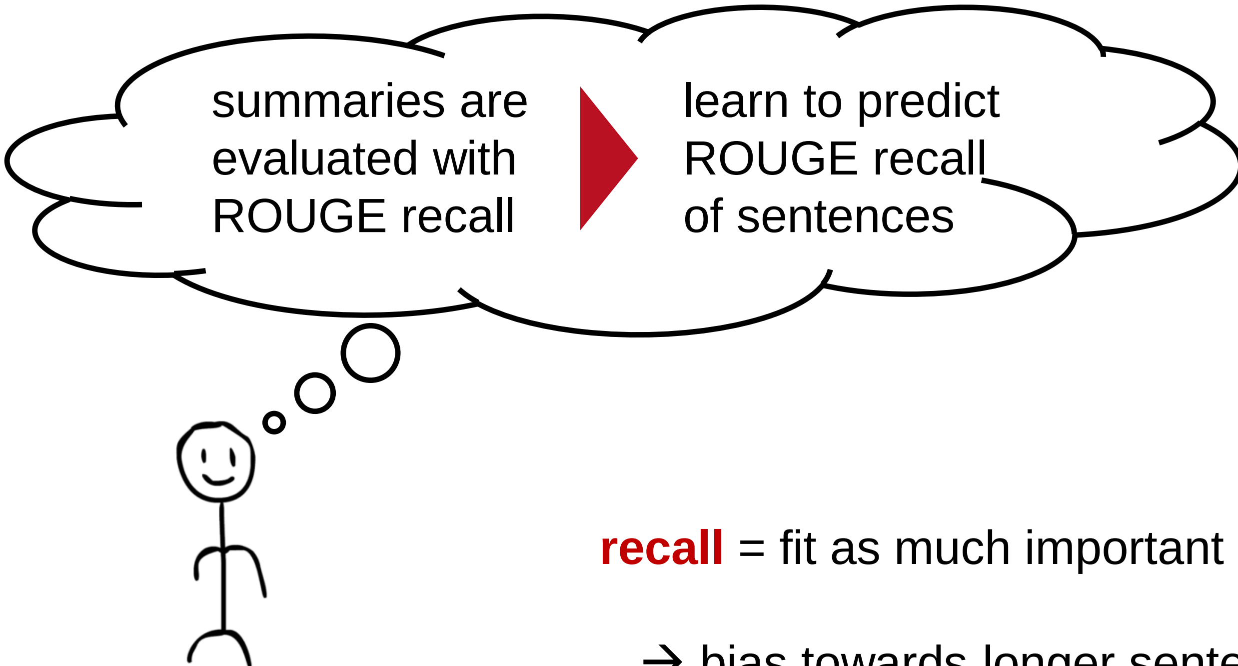
sentence regression



greedy sentence selection

not given in summarization datasets

Which regressand? ROUGE recall?



recall = fit as much important content as possible

→ bias towards longer sentences = wasting space

precision = waste as little space as possible

[Zopf et al., NAACL 2018]

Sentence Length Experiments

regressand	avg. number of sentences		
	DUC 2004	TAC 2008	TAC 2009
random	6.66	5.21	4.89
max. ADW	6.56	5.06	5.11
avg. ADW	5.12	4.13	4.02
R1 Precision	7.76	6.75	6.07
R1 Recall	**3.42**	**2.67**	**2.70**
R2 Precision	7.10	6.13	6.09
R2 Recall	**4.26**	**3.46**	**3.55**

→ selecting according to recall leads to fewer (→ longer) sentences

[Zopf et al., NAACL 2018]

Summarization Experiments

TECHNISCHE UNIVERSITÄT DARMSTADT

regressand	DUC 2004		TAC 2008		TAC 2009	
	R1	R2	R1	R2	R1	R2
random	0.3176	0.0466	0.2958	0.0460	0.2988	0.0463
max. ADW	0.3760	0.1013	0.4255	0.1546	0.3456	0.1105
avg. ADW	0.3850	0.0962	0.4097	0.1243	0.3548	0.0934
R1 Precision	**0.4129**	0.1118	**0.4356**	0.1465	**0.3945**	0.1217
R1 Recall	0.3863	0.0899	0.3928	0.1108	0.3431	0.0837
R2 Precision	0.3918	**0.1273**	0.4346	**0.1819**	0.3781	**0.1364**
R2 Recall	0.3923	0.1207	0.4239	0.1620	0.3742	0.1303

→ greedy selection according to precision leads to high recall in the end

→ prior works predict suboptimal scores

[Zopf et al., NAACL 2018]

Towards Context-free Information Importance Estimation

On the Importance of Information

A Definition of Importance

Evaluating IIE Abilities

The Need for Context-free IIE

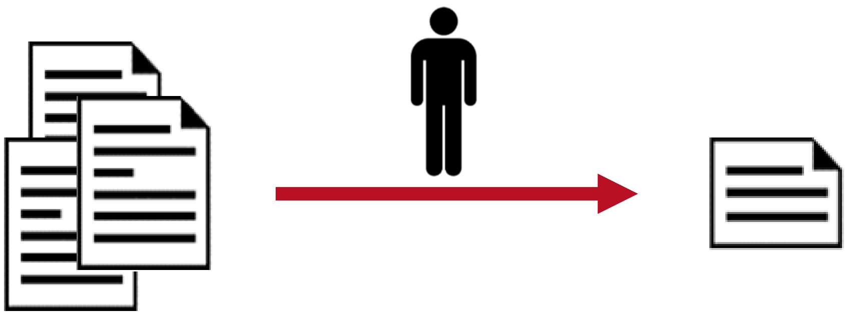
Machine Learning for Context-free Information Importance Estimation

Data

Machine Learning

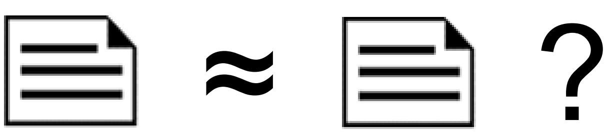
Evaluation

Creating reference summaries is expensive and similarity estimation is unreliable



▶ **expensive**

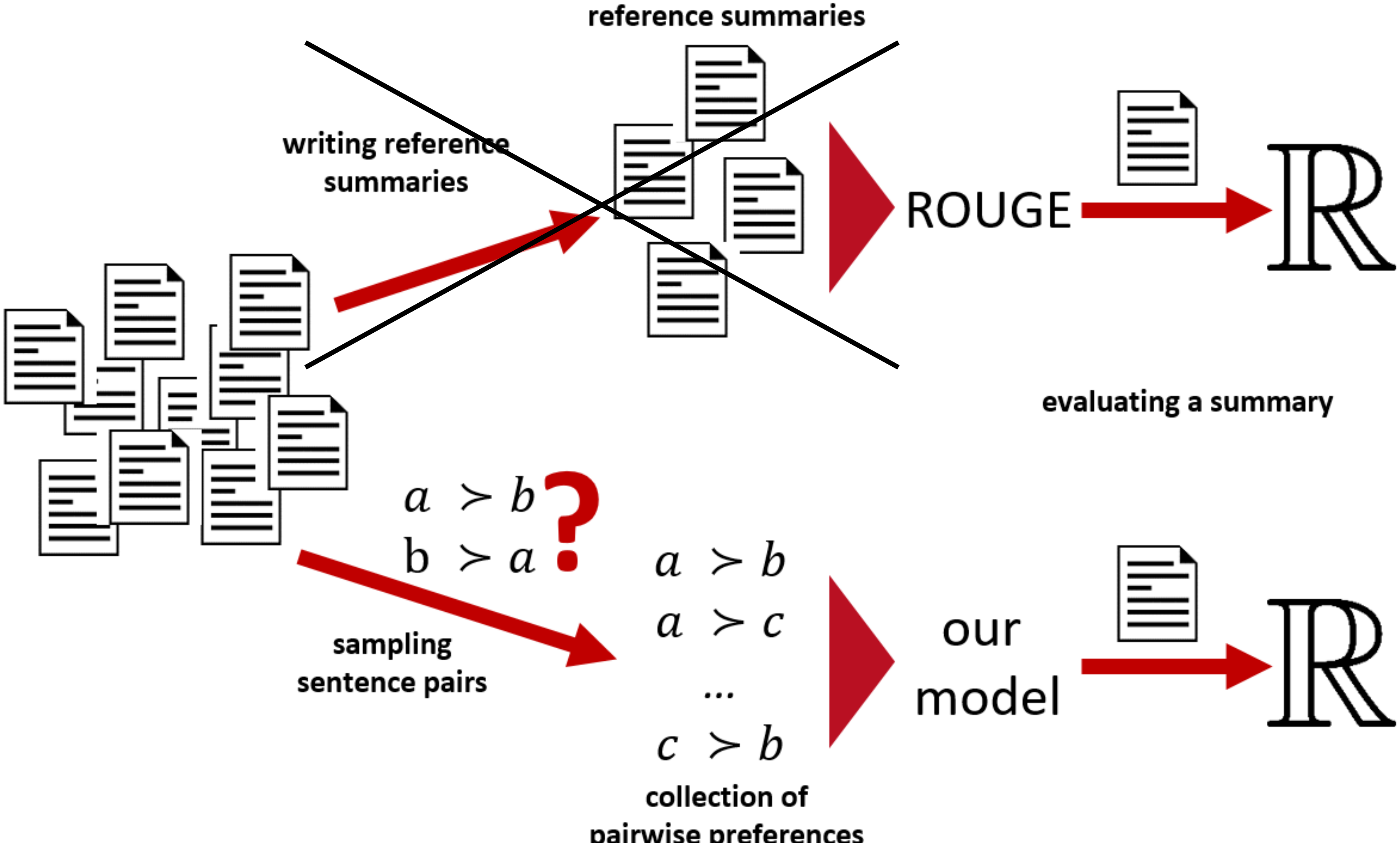
1. reference summaries have to be produced by humans



▶ **unreliable**

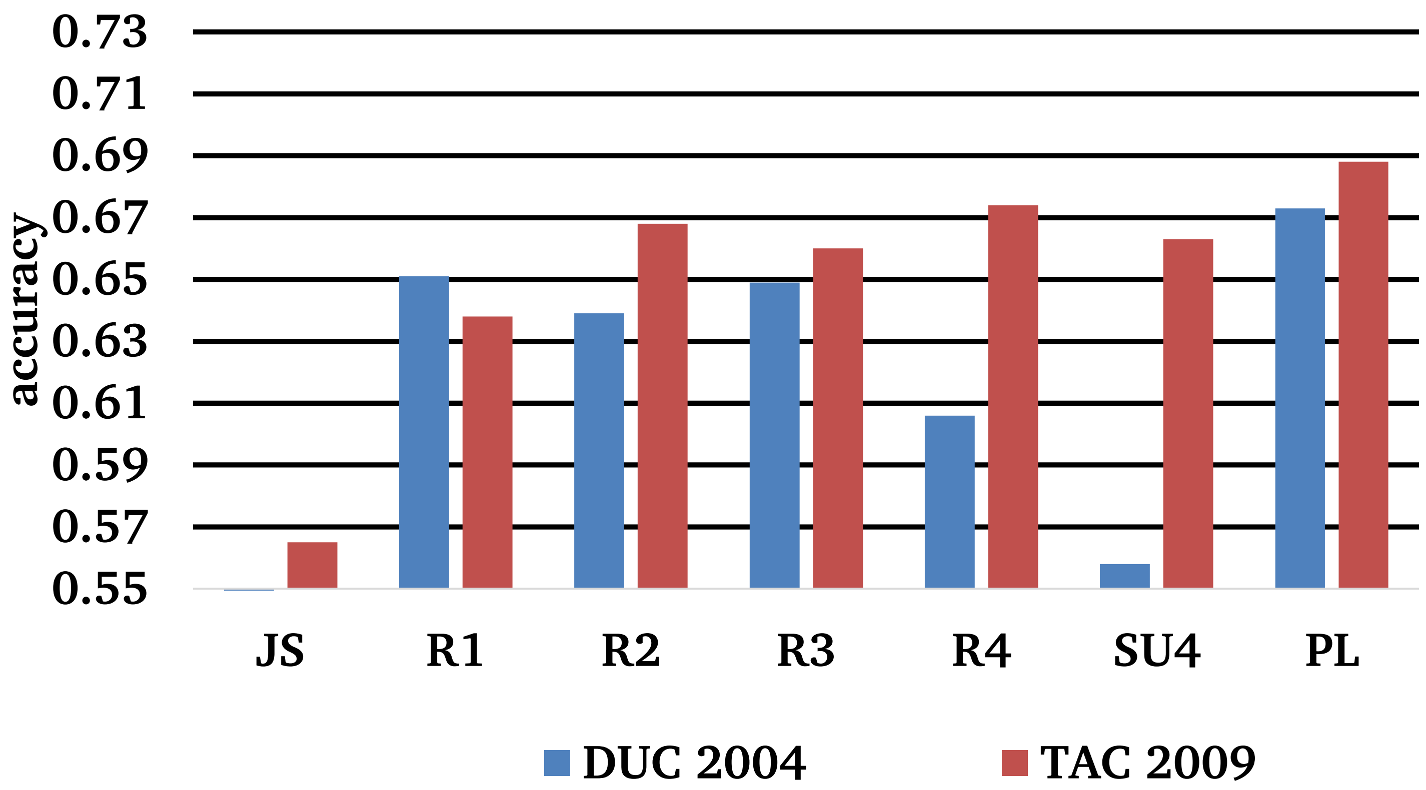
2. computing (semantic) similarity between texts is complex (unsolved problem)

Pairwise preferences over sentences can be used instead



[Zopf, NAACL 2018]

Preference-based evaluation outperforms popular versions of ROUGE

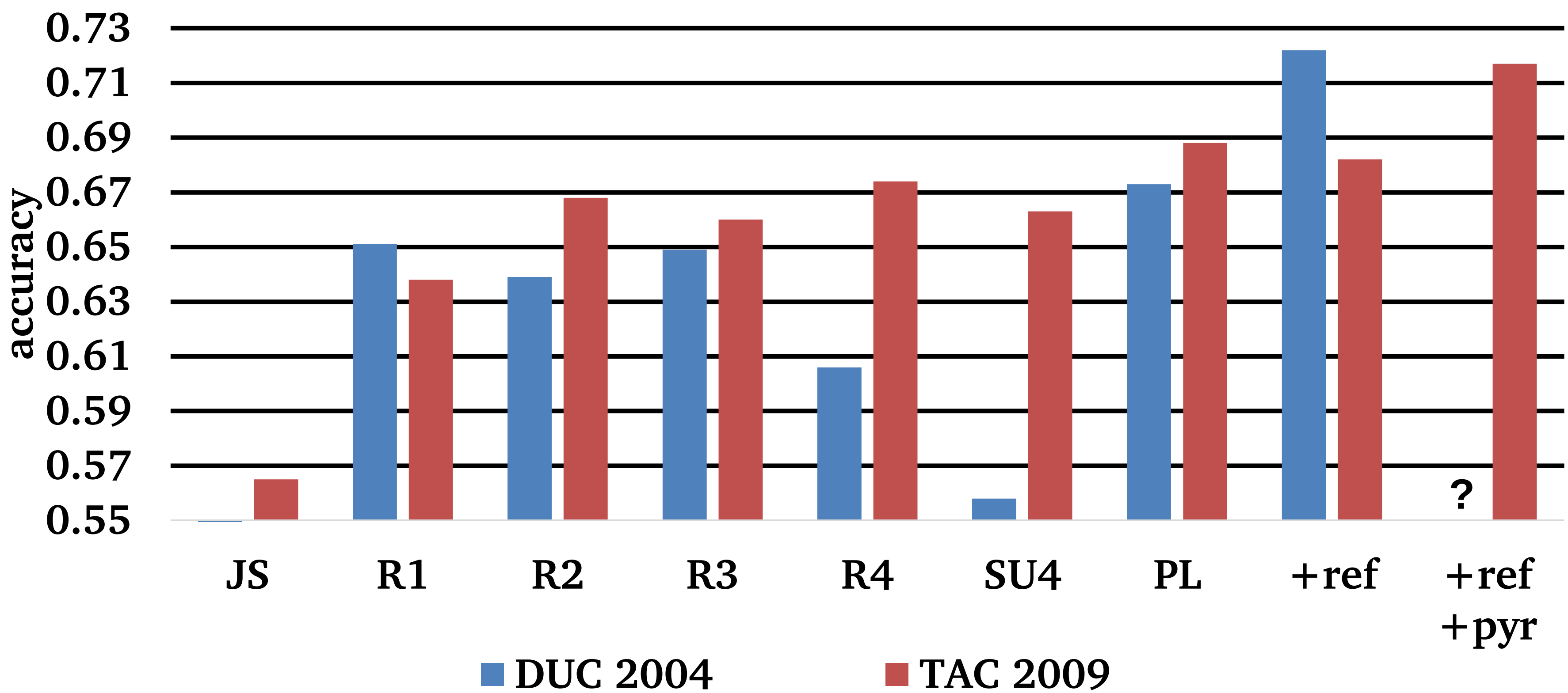


200 preferences per topic

→ ~54 minutes per topic

[Zopf, NAACL 2018]

Automatically generated preferences can further improve performance



+ref = automatic preferences based on reference summaries

+pyr = automatic preferences based on Pyramid scores

[Zopf, NAACL 2018]

Summary: A Definition of Importance

TECHNISCHE UNIVERSITÄT DARMSTADT

Problems

- **vague intuition** of „importance“

  → leads to many discussions based on misunderstandings
- definition of importance only **by example**

Contributions

- two **formal definitions** of information importance

  → common discussion ground
- analysis of **implications**

Future Work

- develop more definitions and compare implications

Knowledge Engineering AIPHES

Summary: Evaluating IIE Abilities

Problems

- summarization **suboptimal** to investigate importance estimation

Contributions

- motivation for importance estimation as **standalone research domain**
  → allows to focus on importance estimation

- three new tasks
  - pairwise **preference** prediction
  - **rankings** according to importance
  - **utility** prediction for information nuggets

Future Work

- new datasets and systems for new tasks

Summary: The Need for Context-free IIE

Problems

- summarization focuses mainly on **newswire documents**
  → special text genre with special properties
- correlation ≠ causality
- correlation can be removed by shuffling / oversampling [Zopf et al., 2016]

Contributions

- **analysis** why summarization systems work the way they do
- proposed **context-free** importance estimation instead

Summary: Data

Problems

- no **heterogeneous** MDS datasets are available
- **large** MDS datasets are missing
  → construction is **too complex and expensive**
- large datasets required for training and testing

Contributions

- novel **cost-effective** corpus construction **approach**
- validated applicability by producing ***h*MDS**
- extended idea to produce **auto-*h*MDS** fully automatically
  → **much larger** than prior MDS datasets

Future Work

- human summarization experiments

Summary: Machine Learning

Contributions

- first **context-free** information importance estimator
  → more (incidental) training data can be used

- resolved a common **misconception** about regressands in sentence regression

Future Work

- explore more context-free information importance estimators

Summary: Evaluation

Problems

- evaluation requires **expensive** reference summaries
- based on rough approximation of **text similarities**

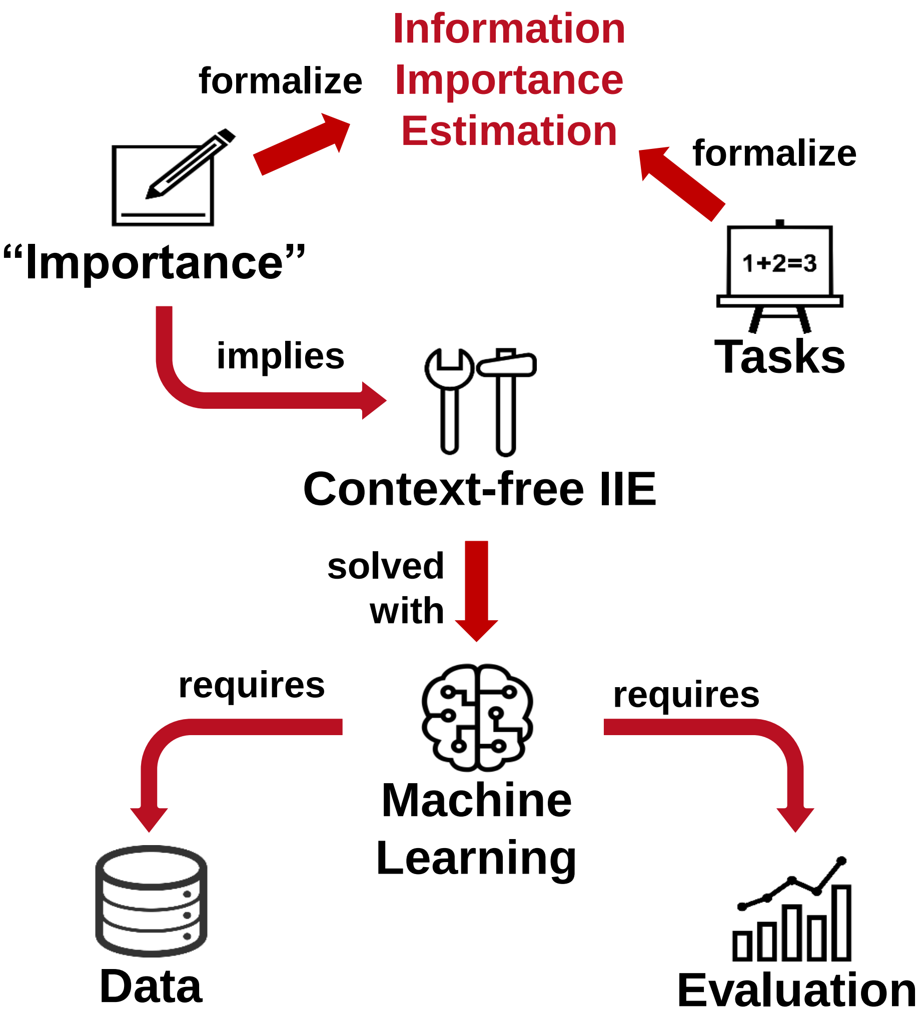
Contributions

- a **new evaluation metric**
  - no reference summaries
  - no text-to-text semantic similarity estimation
- new metric performs better than different versions of ROUGE

Future work

- evaluate on more (diverse) datasets

Towards Context-free Information Importance Estimation



- **Markus Zopf**, et al. What's Important in a Text? An Extensive Evaluation of Linguistic Annotations for Summarization. SNAMS 2018, pp. 272–277
- **Markus Zopf**. Estimating Summary Quality with Pairwise Preferences. NAACL 2018, pp. 1687–1696
- **Markus Zopf**, Eneldo Loza Mencía, and Johannes Fürnkranz. Which Scores to Predict in Sentence Regression for Text Summarization? NAACL 2018, pp. 1782–1791
- **Markus Zopf**. auto-hMDS: Automatic Construction of a Large Heterogeneous Multi-Document Summarization Corpus. LREC 2018, pp. 3228–3233
- **Markus Zopf**, Maxime Peyrard, and Judith Eckle-Kohler. The Next Step for Multi-Document Summarization : A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach. COLING 2016, pp. 1535–1545
- **Markus Zopf**, Eneldo Loza Mencía, and Johannes Fürnkranz. Sequential Clustering and Contextual Importance Measures for Incremental Update Summarization. COLING 2016, pp. 1071–1082
- **Markus Zopf**, Eneldo Loza Mencía, and Johannes Fürnkranz. Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization. CoNLL 2016, pp. 84–94
- **Markus Zopf**. SeqCluSum: Combining Sequential Clustering and Contextual Importance Measuring to Summarize Developing Events over Time. TREC 2015