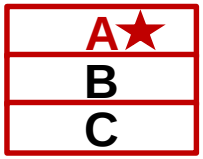
# Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization

**Markus Zopf, Eneldo Loza Mencía and Johannes Fürnkranz**
{zopf@aiphes,eneldo@ke,juffi@ke}.tu-darmstadt.de
https://www.aiphes.tu-darmstadt.de

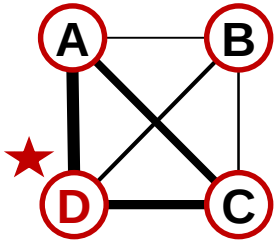
# Importance is estimated by measuring centrality and structural features

## Extractive multi-document summarization nowadays

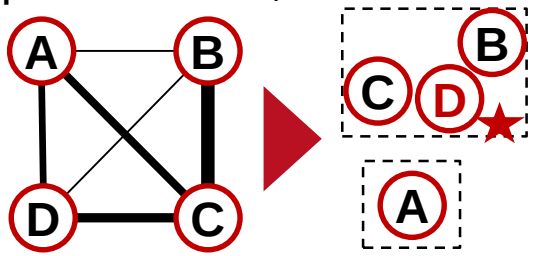
Lead
Sentence Position



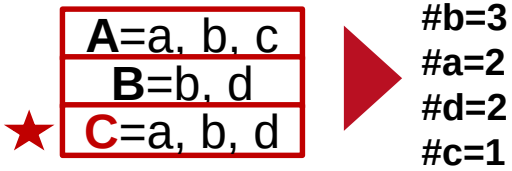

MMR, Carbonell '98 / Submodular, Lin '11
Cosine similarity



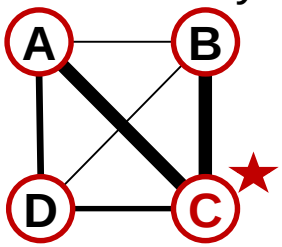

Centroid, Radev '04
topic detection, cluster centroid




ICSI, Gillick '08
number of bi-grams



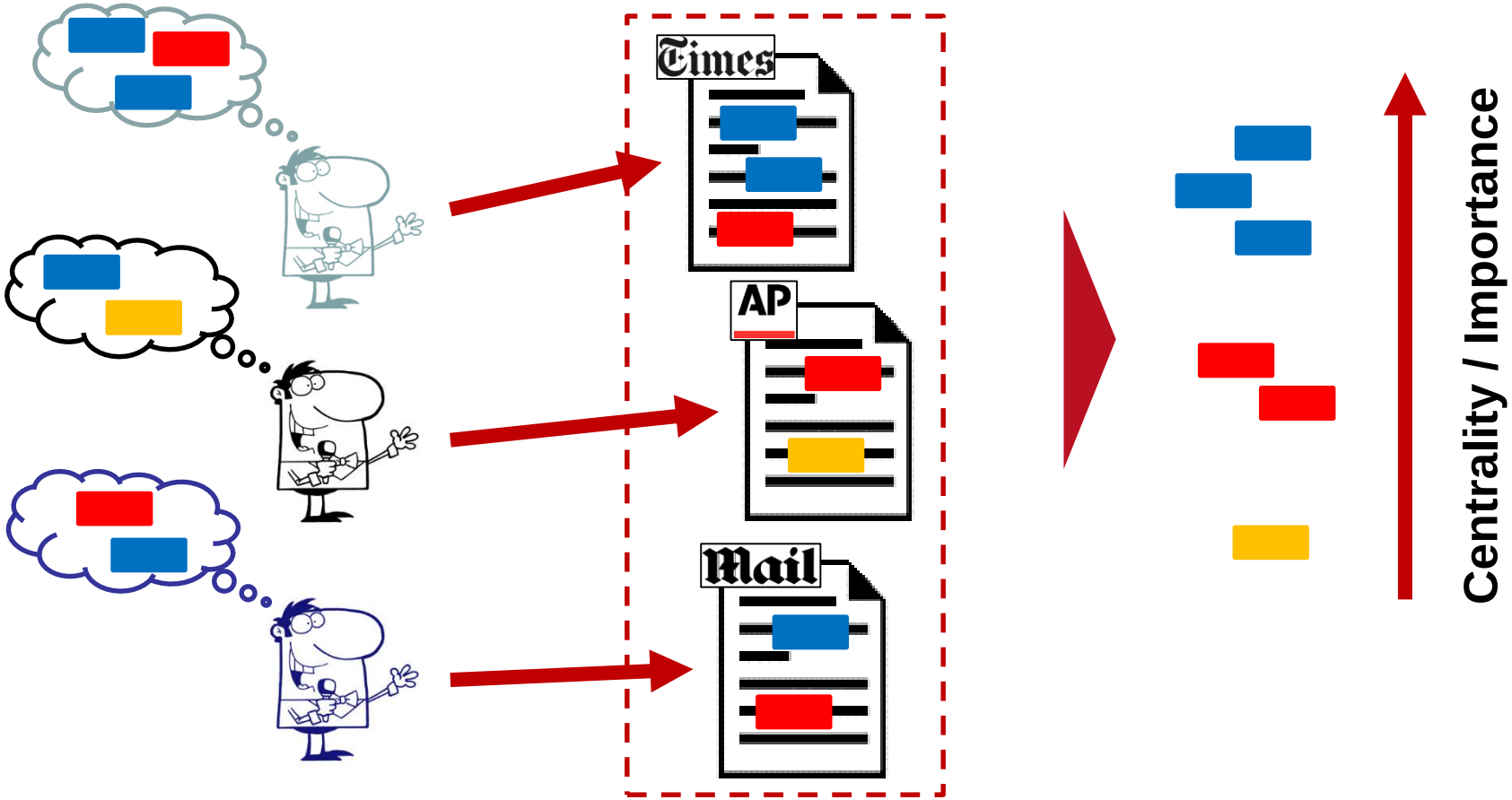

LexRank, Erkan '04
Cosine similarity




★ = best sentence

# Journalists add these features to source documents

## Why are these heuristics used all the time?



# Other text genres do not contain such easy-to-use features

## Summarizing non-newswire documents

film reviews

books

forums

social networks

micro-blogs

blogs

advertisements

debates

video comments

tutorial slides

e-mails

discussions

# Summarize this document by selecting the most important sentence!

**Barack Obama**

Barack Obama graduated from Columbia University and Harvard Law School. Obama is currently serving as the 44th President of the United States. He is a supporter of the Chicago White Sox.

source: Wikipedia

# Summarize this document by selecting the most important sentence!

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

**Barack Obama**

Barack Obama graduated from Columbia University and Harvard Law School. Obama is currently serving as the 44th President of the United States. He is a supporter of the Chicago White Sox.

source: Wikipedia

# Human-like summarization skills require background knowledge

## Why sentence no. 2?

- we **know** that being
  - President of the United States ≻ Columbia University graduate
  - President of the United States ≻ Supporter of the Chicago White Sox
- signal for information importance is not contained in the document
  - centrality
  - sentence position
- summarizing this document requires **background knowledge**

→ **learn from background sources about importance**

# Learn to estimate information importance by investigating background sources

## The CPSum algorithm

background source

e.g. stock market prices,
length of Wikipedia articles,
number of city inhabitants, …

**1. learn about importance**
**2. apply knowledge to summarization**

source documents

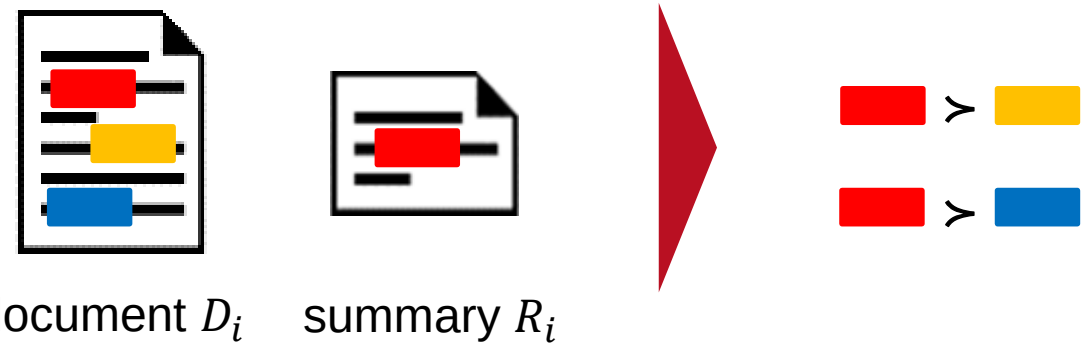
summarization

summary

# Importance is estimated with preference learning by observing promoted objects

## Context-free Importance

**background source:** document-summary pairs, Hermann et. al, 2015




- promoted elements $P_i = D_i \cap R_i = \{$ $\}$
- not promoted elements $N_i = D_i \setminus P_i = \{$ , $\}$

# The more important the objects in a sentence, the more important the sentence

## From observations to sentence scores

**counts:** $\mathrm{n}(a \succ b) = \sum_{(D_i,R_i)\in\mathbb{B}} \mathbf{1}_{P_i}(a) * \mathbf{1}_{N_i}(b)$

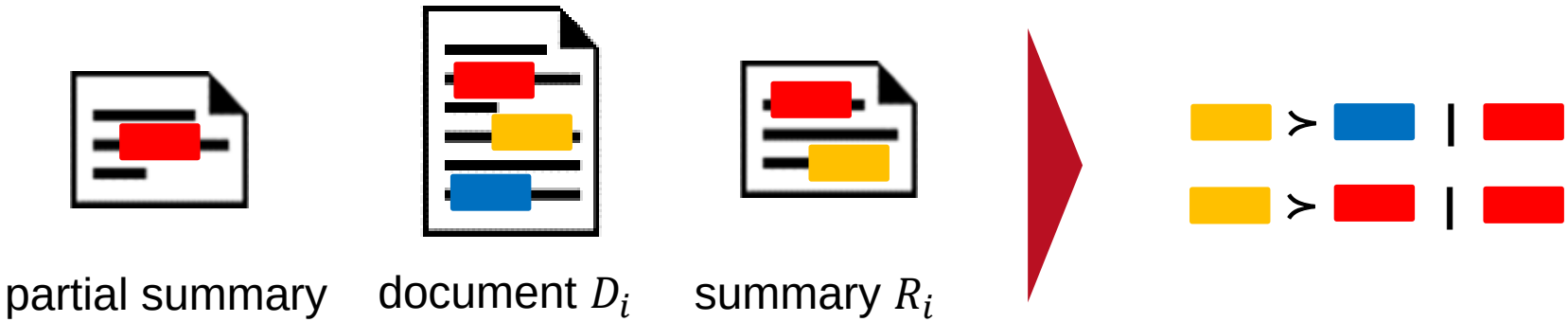
**probability:** $\Pr(a \succ b) = \frac{n\,(a \succ b)}{n\,(a \succ b) + n\,(b \succ a)}$

**object utility:** $\mathrm{v}(o) = \frac{1}{|X|}\sum_{x\in X} \Pr(o \succ x)$

**sentence utility:** $\mathrm{u}(s) = \frac{1}{\mathrm{l}(s)}\sum_{o\in s} v(o)$

# Contextual importance is estimated by observing promoted objects in context

## Contextual Importance

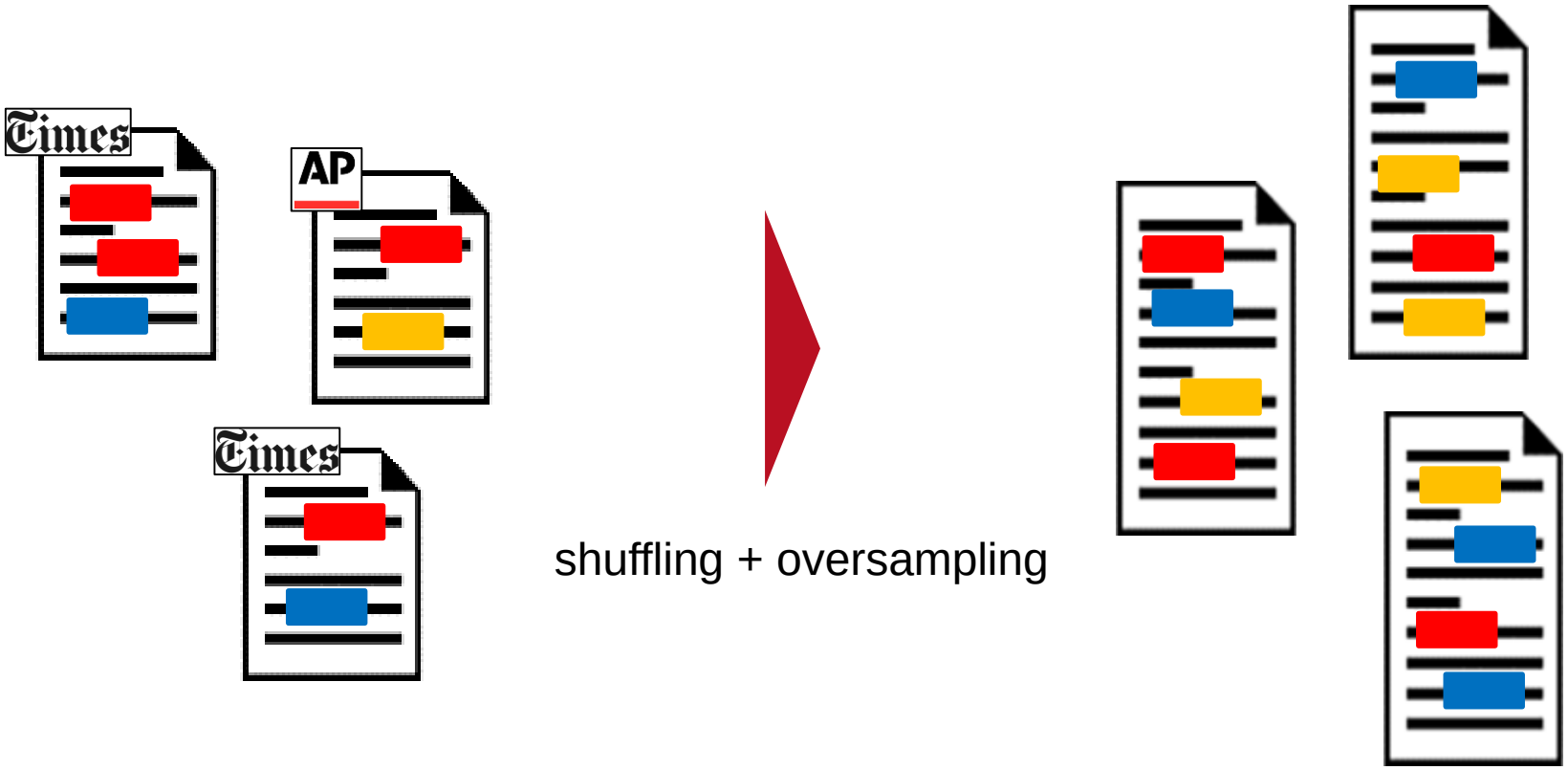



- contextual promoted elements $P_i \mid C = D_i \cap (R_i \setminus \mathrm{C}) = \{$ $\}$
- contextual not promoted elements $N_i \mid C = D_i \setminus (R_i \setminus C) = \{$ , $\}$

- we do not avoid redundancy - we learn which information should be included next

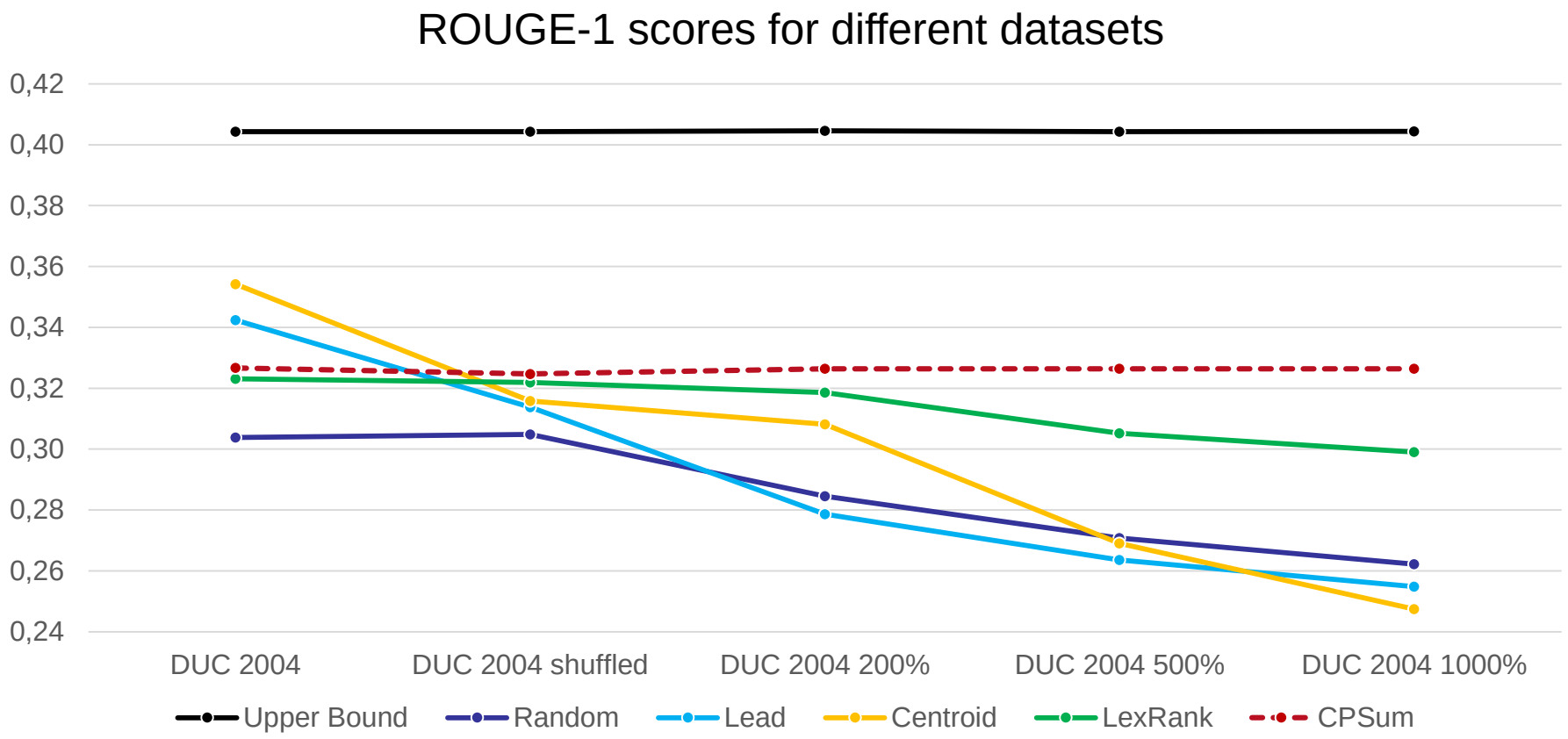
# We hide centrality and sentence position by shuffling and oversampling

## Evaluation: (modified) DUC 2014 MDS dataset



# Performance of all approaches drops whereas CPSum stays constant

## Results



## Summary

- MDS nowadays focusses on centrality and structural features
- newswire data inherently contains easy to exploit features
- not necessarily available in other text genres
- learning human-like importance detection with background knowledge

**Markus Zopf, Eneldo Loza Mencía and Johannes Fürnkranz**
**{zopf@aiphes,eneldo@ke,juffi@ke}.tu-darmstadt.de**
**https://www.aiphes.tu-darmstadt.de**

# Performance of all approaches drops whereas CPSum stays constant

## Results