

Automatische Textzusammenfassung Automatic Text Summarization

Universität zu Köln - Informationsverarbeitung - SoSe 2022

Seminar Computerlinguistik II bei Dr. Jürgen Hermes

Referentinnen: Larissa Rauh, Anne Reiner

Überblick Referat

- Was ist Textzusammenfassung
- Zielgruppe
- Funktionsweise und Methoden
- Beispiel
- Qualitätsmerkmale und Evaluation
- Herausforderungen
- Literatur

Was ist Text- zusammen- fassung?

- Kürzen von Textdokument(en)
- Impuls: steigende Anzahl an Textdaten
- Berücksichtigung sprachlicher Ebenen: Syntax, Semantik, Morphologie

Verortung in der Computer- linguistik

- Teilbereich des textbasierten Information Management (TIM)

TIM Struktur

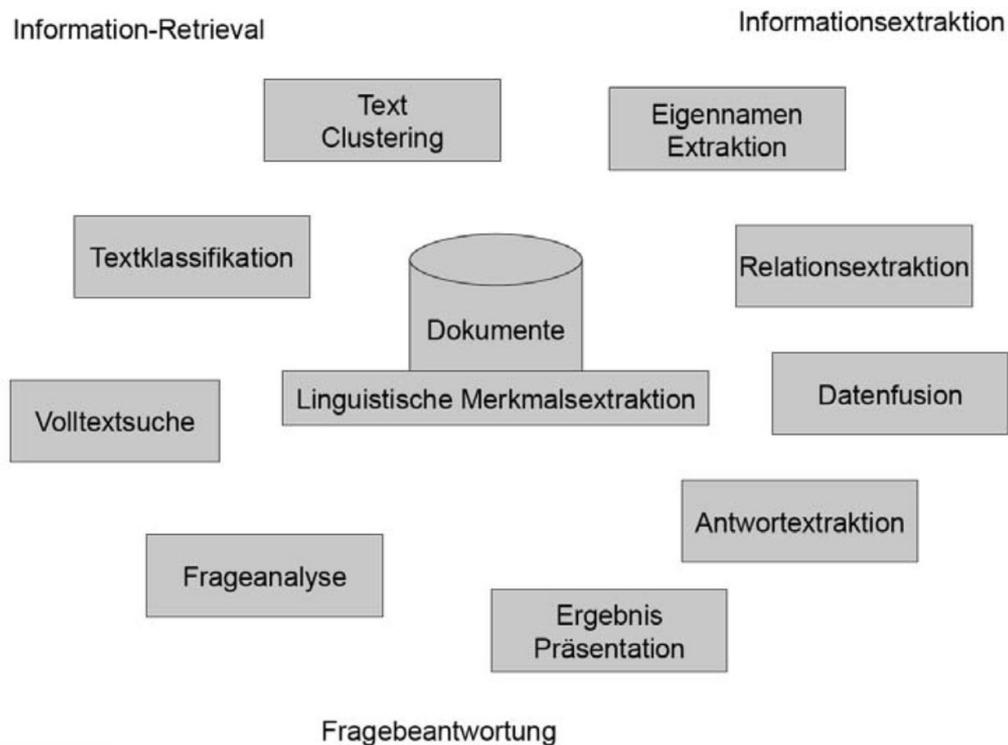


Abbildung 5.1: Text-basiertes Informationsmanagement: Technologieperspektive. Genaue Erläuterungen finden sich im Text.

Wer benötigt Textzusammen- fassungen?

Beispiele..

- Wissenschaftler und Studierende → wiss. Publikationen, Quellen
- Journalisten → Nachrichtenmeldungen
- Händler und Dienstleister → Kundenbewertungen
- Juristen → Urteile
- Personen / Institutionen, die mit längeren und zahlreichen Texten arbeiten

Mehrwert der Anwendungen

- Zeitersparnis
- Kognitive Ersparnis
- Finanzielle Ersparnis
- Überblick
- Cross- bzw. bilinguale Verfahren möglich

Ziele der automatischen TZ

- Ein oder mehrere Textdokumente komprimiert
- Länge angemessen
- Alle relevanten Informationen
- Keine Redundanz
- Strukturiert & Kohärent
- Lesbarkeit / Grammatikalität

Forschungs-
aktivität seit
einigen Jahren
steigend

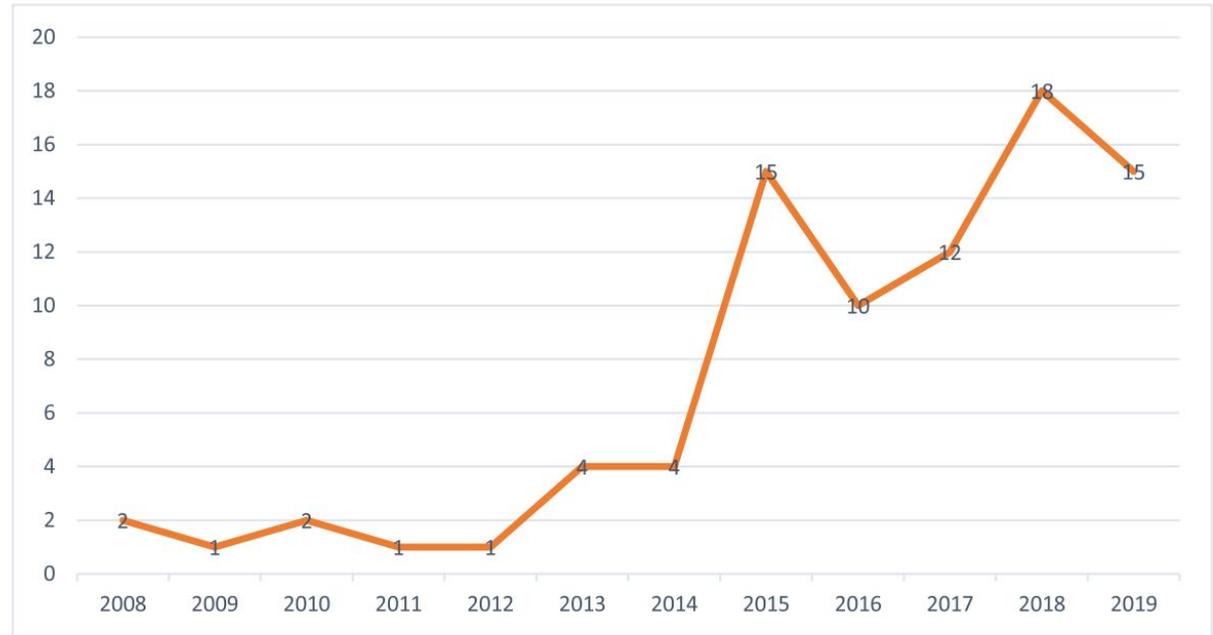


Fig. 4. Distribution over the past ten years for selected studies.

Extraktion & Abstraktion

Extraktive Verfahren

- Relevante Sätze aus dem Originaltext
- Grammatik ist immer korrekt
- Zentrale Themen abdecken
- Kohäsion erhalten
- inzwischen recht ausgereift

Abstraktive Verfahren

- Paraphrasieren
- Annäherung an menschliche Praxis: zuerst semantische Repräsentation, dann Zusammenfassung in eigenen Worten in eigener Reihenfolge
- Programm muss Hintergrundwissen integrieren

Forschungs- themen 2009-2018

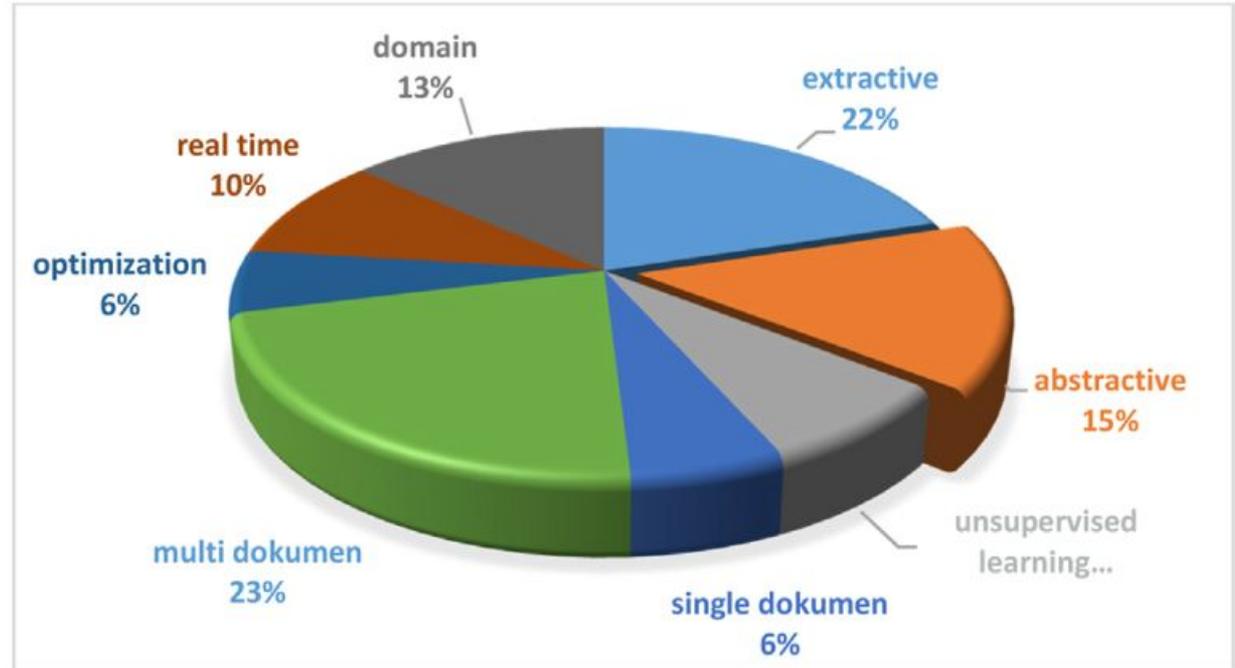


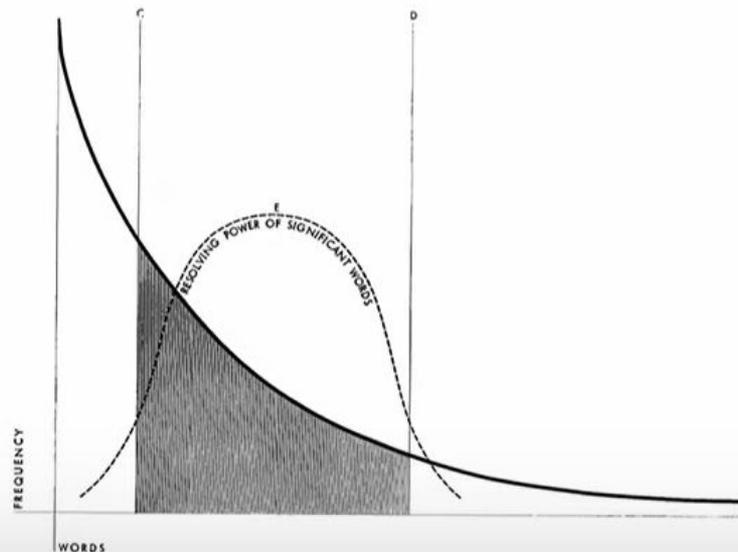
Fig. 8. Distribution of Topic or Trend Research.

Methoden

Luhn's Method (1958)

- Data Pre-Processing
 - Stop words
 - Stemming
 - Tokenizing
- Frequenz signifikanter Phrasen

Figure 1 **Word-frequency diagram.**
Abscissa represents individual words arranged in order of frequency.



Methoden

Edmundson's Method (1968)

- Cue words
 - bonus words ↑
 - stigma words ↓
 - null words –

→ extraktiv

Methoden

FRUMP (1979)

- Fast Reading **U**nderstanding and **M**emory
Programm
- Templates

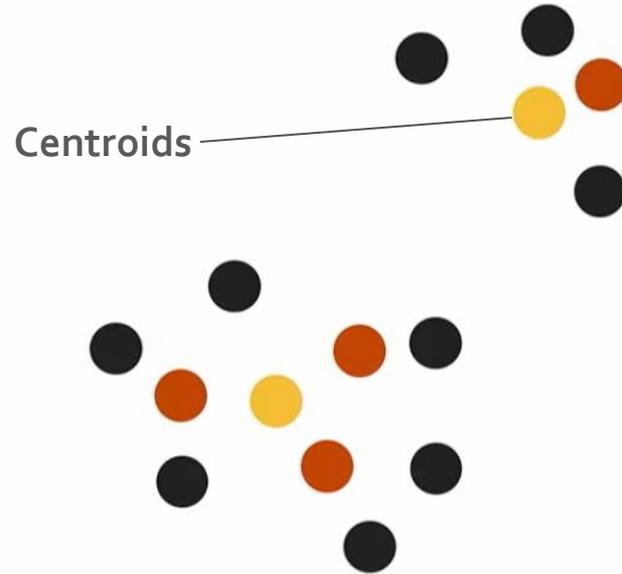
→ abstraktiv

Methoden

Maximal
Marginal
Relevance

Centroid-based Method

- Textclustering



Methoden

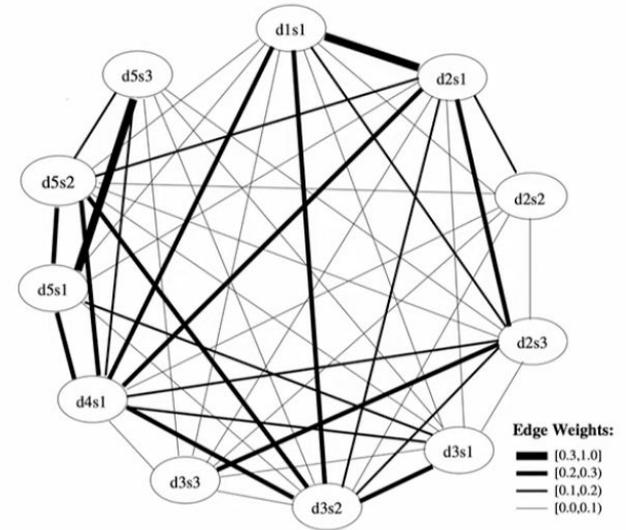
Maximal
Marginal
Relevance

LexRank

- Graphen-basierte Methode
- Ähnlichkeitsmatrix (Similarity Matrix)
- multi-document summary

TextRank

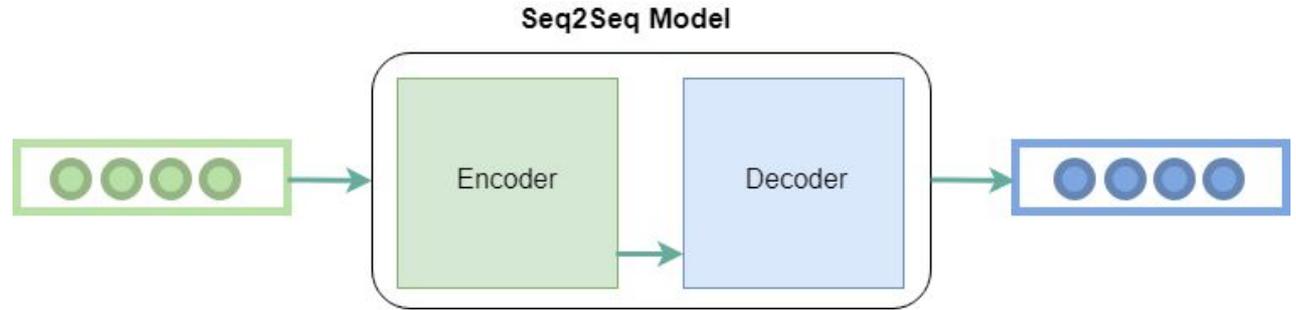
- single-document summary



Methoden

seq2seq

Sequence to Sequence



- Entwickelt von Google für die maschinelle Übersetzung
- Verwendet Deep Learning → Berücksichtigung der vorausgehenden Wortumgebung
- Encoder-Decoder-Netzwerk

Evaluations- methoden



- Qualität der Textzusammenfassung?

Manuelle Zusammen- fassungen

durch
Menschen

- Unterschiedliche Längen der Zus.fassungen
- Thematische Akzente
- Vorwissen, Kontext und Absichten werden integriert
- Mensch wählt aus, was für ihn / sie bedeutungsvoll ist
- Auch ein Mensch erreicht keine hundertprozentige Vollständigkeit und Präzision

Erinnerung:

Ziele der
automatischen
TZ

- Ein oder mehrere Textdokumente komprimiert
- Länge angemessen
- Alle relevanten Informationen
- Keine Redundanz
- Strukturiert & Kohärent
- Lesbarkeit / Grammatikalität

Evaluations- methoden

Task based

- Zwei Teilnehmergruppen
- erste Gruppe erhält Originaldokument
- zweite Gruppe erhält maschinelle TZ
- beide Gruppen bekommen Aufgabe

Treffen Personen auf Basis der TZ dieselbe Entscheidung wie Personen, die das Originaldokument haben?

Comparing

- Vergleich Automatische TZ / Manuelle TZ
- Manuelle TZ hier Goldstandard, Mensch als Maßstab
- **Etablierte Methode: Precision & Recall**

Precision & Recall

- Vergleich manuelle / automatische TZ
- Verfahren für Evaluation extraktiver TZ

	manually selected	manually not selected
automatic selected	True positives	False positives
automatic not selected	False negatives	True negatives

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Precision & Recall

- Vergleich manuelle / automatische TZ
- Verfahren für Evaluation extraktiver TZ

	manually selected	manually not selected
automatic selected	True positives	False positives
automatic not selected	False negatives	True negatives

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Utility

Problem bei Precision & Recall:

- Wenn mehrere Sätze ähnliche Information enthalten, treffen Mensch und Maschine womöglich nicht die selbe Auswahl
- Trotzdem ist die Auswahl durch den Menschen nicht generell relevanter (objektiv gesehen)
- Lösung: Ranking des Relevanzgrades der in Frage kommenden Sätze von 1-10 durch den Menschen
-> Summarizer erhält Punkte für ausgewählten Satz

ROUGE-N & ROUGE-L

Recall-Oriented-Understudy for Gisting Evaluation

- Hoher Recall-Wert = hoher ROUGE N-Wert (Recall: Wurden alle relevanten Sätze gefunden?)
- automatische TZ wird mit Set manueller Textzusammenfassungen (verschiedener Menschen) verglichen
- Hierbei Nutzung der N-Gramm Überschneidungen, kommen sie sowohl in (mehr als) einer manuellen als auch in der automatischen TZ vor?
- ROUGE-L: Suche nach längster gemeinsamer Sequenz

Beispiel für
Tool

<https://smmry.com>

(ungelöste) Probleme

- Mehrdeutigkeit (auf Satz- und Wortebene)
- Redundanz
- Synonyme, Neologismen, implizite Nennungen
- Niedrigfrequentes wird aussortiert → neue Aspekte bleiben unerfasst
- Plagiatsgefahr, wenn Zusammenfassung rein extraktiv
- Maßstab Mensch nicht eindeutig

Literatur

Baxendale, P. B. (1958). Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, 2(4), 354–361.

Carbinell, J. & Goldstein, J. (2017). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *ACM SIGIR Forum*, 51(2), 209–210.

Carstensen, K., Ebert, C., Ebert, C., Jekat, S., Langer, H. & Klabunde, R. (2009). *Computerlinguistik und Sprachtechnologie: Eine Einführung* (3. überarb. u. erw. Aufl. 2010 Aufl.). Spektrum Akademischer Verlag.

DeJong, G. (1979). Prediction and Substantiation: A New Approach to Natural Language Processing*. *Cognitive Science*, 3(3), 251–273.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, 16(2), 264–285.

Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization. *Journal of Artificial Intelligence Research*, 22, 457–479.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2), 159–165.

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A. & Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4), 1029–1046.