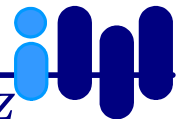


EXCERPT

Ein integriertes

---

Within-Document-Retrieval-  
und Summarizing-System



# Einführung

---

## ■ Zielsetzung:

- Integriertes System zum Within-Document-Retrieval und Summarizing:
  - 'Suchmaschine' für Passagen in Texten
  - einheitlicher Algorithmus, verminderter Implementierungsaufwand
- Techniken des Summarizing auf Within-Document-Retrieval übertragen:
  - Bewertung von Passagen nach bestimmten Kriterien
  - Auswahl von Information gemäß Nutzerinteresse

# Status Quo

---

## ■ Within-Document-Retrieval-Systeme:

- ProfileSkim [Harper & al. 2004]
- Passagen-Retrieval-Systeme (z. B. [Salton & al. 1993])
- Textdatei-Suchsysteme (z. B. SearchWithin)

## ■ Summarizing-Systeme:

- Überblick in [Mani & Maybury 1999]
- Online-/Offline-Systeme
  - Summarizer des Language Technologies Research Centers (<http://search.iiit.net/~jags/summarizer/index.cgi>)
  - Open Text Summarizer (<http://libots.sourceforge.net/>)

# Summarizing

---

## ■ Summaries:

### ● Unterscheidungen:

- extractiv versus abstraktiv
- indikativ versus informativ
- generisch versus spezifisch (nutzer-/themazentriert)

### ● Verfahren:

- korpusbasiert (statistisch, trainierbar)
- diskursorientiert (z. B. lexikalische Ketten)
- wissensbasiert (Domänenmodellierung)

# Summarizing

---

- Gebräuchliche Parameter zur automatischen Ermittlung relevanter Summary-Passagen:
  - Satzposition (text/absatz-initial/final)
  - Kohäsion/Kohärenz (Termrelationen, Diskursstruktur)
  - Titel-/Schlüsselterm (Bonus-/Maluswörter)
  - Topikalität/Spezifität (Termfrequenz)
  - Wortkategorien (Funktions-/Inhaltswörter)
  - Satzlänge (z. B. mehr als 5 Wörter)
  - ...

# Summarizing

---

- Speziell Wortkategorien (vgl. [Edmundson 1969]; [Mittal & al. 1999]; [Banko & al. 1999]; [Goldstein & al. 1999]):

- positiv:

- semantisch relationierte Wörter
- Komparative/Superlative
- konklusive Adverbien
- indefinite Artikel
- Eigennamen

- negativ:

- Pronomen (demonstrative, personale 1. Person)
- Negationen
- Auxiliare
- Konjunktionen
- Präpositionen
- Honorifika

# EXCERPT

---

## ■ Ansatz/Idee hinter EXCERPT:

- Ausgangspunkt topikfokussierte Echtzeit-Summary:
  - 0 Suchterme: generisches Summary
  - 1–N Suchterme: spezifisches Summary
- variable Einstellungen durch Nutzer:
  - Ausgabe-Einheit: Sätze oder Paragraphen
  - Ausgabe-Größe: relative oder absolute Anzahl Einheiten
  - Ausgabe-Weise: (dis)kontinuierliche Einheiten (mit/ohne Lücken)
  - scharfe/unscharfe Suche: direkte/indirekte 'Matches' (precision-/recall-orientiert)

# mammoth elephant

Evaluation Extraction

Analysis BonusMalus

2 passages found:

Rank=0 Score=6 Items=8  
 [0] Regenerating a Mammoth for \$10 Million [5] There are talks on how to modify the DNA in an elephant's egg so that after each round of changes it would progressively resemble the DNA in a mammoth egg. [6] The final-stage egg could then be brought to term in an elephant mother, and mammoths might once again roam the Siberian steppes. [9] Mammoths, ice-age relatives of the elephant, were hunted by the modern humans who first learned to inhabit Siberia some 22,000 years ago. [10] The mammoths fell extinct in both their Siberian and North American homelands toward the end of the last ice age, some 10,000 years ago. [12] They have already been able to calculate that the mammoth's genes differ at some 400,000 sites on its genome from that of the African elephant. [13] There is no present way to synthesize a genome-size chunk of mammoth DNA, let alone to develop it into a whole animal. [14] But Dr. Schuster said a shortcut would be to modify the genome of an elephant's cell at the 400,000 or more sites necessary to make it resemble a mammoth's genome.

Rank=1 Score=1.5 Items=6  
 [1] Scientists are talking for the first time about the old idea of resurrecting extinct species as if this staple of science fiction is a realistic possibility, saying that a living mammoth could perhaps be regenerated for as little as \$10 million. [8] A scientific team headed by Stephan C. Schuster and Webb Miller at Pennsylvania State University reports in Thursday's issue of Nature that it has recovered a large fraction of the mammoth genome from clumps of mammoth hair. [11] Dr. Schuster and Dr. Miller said there was no technical obstacle to decoding the full mammoth genome, which they believe could be achieved for a further \$2 million. [15] The cell could be converted into an embryo and brought to term by an elephant, a project he estimated would cost some \$10 million. [17] There have been several Russian attempts to cultivate eggs from frozen mammoths that look so perfectly preserved in ice. [29] Rudolph Jaenisch, a biologist at the Whitehead Institute in Cambridge, said the proposal to resurrect a mammoth was "a wishful-thinking experiment with no realistic chance for success."

Bonusterms	Malusterms
aftermost	alas
backmost	apropos
best	by the bye
best-known	by the way
best-selling	hereby
better	herewith
bigger	incidentally
biggest	
blacker	
blackest	
bluer	
bluest	
bottommost	
brand-new	
brand-newly	
browner	
brownest	
cleaner	
cleanest	
considerable	
considerably	
critical	
critically	
crucial	
crucially	
dangerous	
dangerously	
decisive	
decisively	

Load Save Load Save

Passages	Accuracy	Topicality	Significancy	Novelty	Informationality	Discursivity	Specificity
Unit: Sentences		Topicality: 1.00		Centrality: 0.00		Themacity: 1.00	
Size: 25%							
Span: 0							

Read Write

Evaluation Extraction

4 passages found:

Rank=0 Score=34.66666666 Items=8

[1] Scientists are talking for the first time about the old idea of resurrecting extinct species as if this staple of science fiction is a realistic possibility, saying that a living mammoth could perhaps be regenerated for as little as \$10 million. [3] Though the stuffed animals in natural history museums are not likely to burst into life again, these old collections are full of items that may contain ancient DNA that can be decoded by the new generation of DNA sequencing machines. [4] If the genome of an extinct species can be reconstructed, biologists can work out the exact DNA differences with the genome of its nearest living relative. [8] A scientific team headed by Stephan C. Schuster and Webb Miller at Pennsylvania State University reports in Thursday's issue of Nature that it has recovered a large fraction of the mammoth genome from clumps of mammoth hair. [9] Mammoths, ice-age relatives of the elephant, were hunted by the modern humans who first learned to inhabit Siberia some 22,000 years ago. [14] But Dr. Schuster said a shortcut would be to modify the genome of an elephant's cell at the 400,000 or more sites necessary to make it resemble a mammoth's genome. [26] A third issue is that the DNA of living cells can be modified only very laboriously and usually at one site at a time. [27] Dr. Schuster said he had been in discussion with George Church, a well-known genome technologist at Harvard Medical School, about a new method Dr. Church has invented for modifying some 50,000 genomic sites at a time.

Rank=1 Score=30.36666668 Items=8

[2] The same technology could be applied to any other extinct species from which one can obtain hair, horn, hooves, fur or feathers, and which went extinct within the last 60,000 years, the effective age limit for DNA. [5] There are talks on how to modify the DNA in an elephant's egg so that after each round of changes it would progressively resemble the DNA in a mammoth egg. [7] The same would be technically possible with Neanderthals, whose full genome is expected to be recovered shortly, but there would be several ethical issues in modifying modern human DNA to that of another human species. [11] Dr. Schuster and Dr. Miller said there was no technical obstacle to decoding the full mammoth genome, which they believe could be achieved for a further \$2 million. [13] There is no present way to synthesize a genome-size chunk of mammoth DNA, let alone to develop it into a whole animal. [17] There have been several Russian attempts to cultivate eggs from frozen mammoths that look so perfectly preserved in ice. [25] Dr. Schuster has found that hair is a much purer source of the host's DNA, with the keratin serving to seal it in and largely exclude bacteria. [28] The method has not yet been published, and until other scientists can assess it they are likely to view genome engineering on such a scale as being implausible

Analysis BonusMalus

Bonustersms	Malustersms
aftermost	alas
backmost	apropos
best	by the bye
best-known	by the way
best-selling	hereby
better	herewith
bigger	incidentally
biggest	
blacker	
blackest	
bluer	
bluest	
bottommost	
brand-new	
brand-newly	
browner	
brownest	
cleaner	
cleanest	
considerable	
considerably	
critical	
critically	
crucial	
crucially	
dangerous	
dangerously	
decisive	
decisively	

Load Save Load Save

Passages Accuracy Topicality Significancy Novelty Informationality Discursivity Specificity

Unit: Sentences

Size: 25%

Span: 0

Topicality: 1.00

Centrality: 0.00

Themacity: 1.00

Read Write

# EXCERPT

---

## ■ Funktionsweise von EXCERPT:

### ● Analysephase:

- Satzerkennung

- Term-Normalisierung

- Mehrwortterm-Erkennung

- Indexerstellung:

- Term-Satz-Index (Retrieval)

- Term-Term-Index (Assoziationsindex)

# EXCERPT

---

- Bewertungsphase (Ermittlung von Parametern in 6 Kategorien):
  - Kohärenzstruktur (lexikalisch-semantische Relationen gemäß WordNet) unter Einbeziehung der
    - Frequenz von Termen in Text/Sprache (Termspezifität)
    - Distanz zwischen Termen (textuell, lexikalisch)
    - Termposition im Satz (Topikalität)
  - Bonus-/Maluswörter (u. a. Komparative/Superlative)
  - Paragraphen-Initialität, Term-Innovativität
  - Deklarativität (Aussagesätze gegenüber Fragesätzen bevorzugt)
  - Indefinite Determinatoren, Personalpronomen (1. Person)
  - Zahlen, Eigennamen

# EXCERPT

---

## ■ Bewertung der Sätze:

1. Die Parameter jeder Kategorie werden pro Satz verrechnet (z. B. +Bonuswörter–Maluswörter).
  2. Für jede Kategorie wird dann eine eigene Rankingliste der so ermittelten Scores erzeugt.
  3. Die 6 Ranking-Indexe mit den Scores werden pro Satz verrechnet → Ermittlung der durchschnittlichen Ranking-Position pro Satz.
- Für jeden Satz wird ein retrieval-unabhängiger Gesamtscore berechnet, der eine Aussage über die 'Relevanz' (Wichtigkeit, Zentralität) eines Satzes macht

# EXCERPT

---

## ■ Retrieval:

1. Analyse der Eingabeterme wie Text (Mehrwortterm-Erkennung etc.)
2. Selektion derjenigen Sätze, die Suchterme (in)direkt enthalten:
  - precision-orientiert: nur gleiche (normalisierte) Terme
  - recall-orientiert: auch semantisch relationierte Terme (Synonyme, Hyp[er]onyme, Antonyme usw.)
3. Sortierung der Sätze nach Retrievalscores und Satzscores:
  1. Anzahl gefundene verschiedener Suchterme (AND → OR)  
Anzahl gefundene gleicher Suchterme  
semantische Distanz Suchterm–Satzterm und Termtopikalität
  2. Satzscores (aus retrieval-unabhängiger Bewertung zuvor)
4. Clusterung der selektierten Sätze zu Ausgabe-Einheiten und gerankte Ausgabe der Cluster

# EXCERPT

Selektion von Sätzen aus Text T mit M Sätzen anhand von N Suchtermen:  
Wenn  $N = 0$ , dann alle M Sätze selektiert (Summary-Modus)  
Wenn  $N > 0$ , dann  $0-M$  Sätze selektiert (Retrieval-Modus)

## Text T mit Suchterm S

Satz [0] = Score 1.0

Satz [1] = Score 1.5

Satz [2] = Score 2.0

Satz [3] = Score 3.0

Satz [4] = Score 2.5

Satz [5] = Score 2.7

Satz [6] = Score 3.5

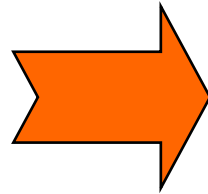
Satz [7] = Score 1.5

Satz [8] = Score 0.5

Satz [9] = Score 1.0

...

Satz [M-1] = Score ...



## Clustering & Ranking

Passagen-Parameter:

- Ausgabeeinheit: Sätze
- Ausgabegröße: 3 Sätze (maximal)
- Ausgabeweise: diskontinuierlich mit Spanne = 2

Ergebnis-Ranking:

Cluster 0 = Score 8.7 aus Sätzen [4,5,6]

Cluster 1 = Score 3.0 aus Sätzen [0,2]

Cluster 2 = Score 1.0 aus Sätzen [9]

...

# EXCERPT

---

## Clustering-Algorithmus:

1. Probiere sukzessive jeden Satz als Pivot-Satz und konstruiere daraus eine initiale Einsatz-Passage, an die weitere Sätze angefügt werden können;
2. wähle einen anderen Satz mit dem höchsten Score, der sich innerhalb der Spanne um den ersten oder letzten Satz der aktuellen Passage befindet;
3. füge diesen Satz in die aktuelle Passage ein, d. h. erweitere die Grenzen der Passage nach oben oder unten;
4. wenn die Mindestgröße der Passage noch nicht erreicht ist, gehe zu Schritt 2; ansonsten gehe zu Schritt 5;
5. wenn die aktuell konstruierte Passage einen höheren Gesamtscore aufweist als die bisher beste Passage, vermerke die aktuelle Passage als beste;
6. wenn noch nicht alle Sätze als Pivotsatz probiert wurden, gehe zu Schritt 1; ansonsten gehe zu Schritt 7;
7. markiere alle Sätze der ermittelten besten Passage als benutzt (verhindere, dass diese weiter als Pivot- oder Anlagerungssatz verwendet werden);
8. gib die beste Passage aus (oder speichere sie zwischen) und konstruiere die nächste Passage, sofern noch Sätze übrig sind.

# EXCERPT

---

- Performance-Messungen anhand manuell bewerteter Texte (noch in Arbeit):
  - Summarizing (vor allem zur Kalibrierung des Systems):
    - 6 Nachrichten-Texte von [Zechner 1995] aus Daily Telegraph Corpus: von 13 Personen bewertet, relativ hohe Übereinstimmungen zwischen Bewertern
    - 7 Eigene Texte
    - (New York Times)
  - Within-Document-Retrieval:
    - 7 Eigene Texte
    - ?

# EXCERPT

---

## ■ Recall/Precision-Maße:

- generell Summarizing/Within-Document-Retrieval:
  - Recall = ermittelte relevante Sätze / relevante Sätze
  - Precision = ermittelte relevante Sätze / ermittelte Sätze
- speziell Summaries [Mani 2001: 230]:
  - Recall = ermittelte relevante Sätze / Länge des manuellen Referenzsummaries (relevante Sätze)
  - Precision = ermittelte relevante Sätze / Länge des maschinellen Summaries (ermittelte Sätze)
  - ➔ da EXCERPT grundsätzlich alle Sätze gerankt ausgibt, werden nur die M Sätze des Referenzsummaries betrachtet; d. h. Anzahl ermittelter = Anzahl relevanter Sätze (Recall = Precision)

# EXCERPT

---

## ■ Performance Zechner-Texte:

<u>Text</u>	<u>BL-Initial</u>	<u>Zechner</u>	<u>EXCERPT</u>
A	0.33	0.50	0.67
B	0.50	0.33	0.67
C	0.43	0.57	0.71
D	0.33	0.67	0.67
E	0.29	0.43	0.29
F	0.67	0.67	0.67
Ø	0.43	0.53	0.61

Hinweis: Zechners System arbeitet mit positionalem Kriterium

# EXCERPT

---

- Auswertung eigener Texte:
  - 14 Dokumente unterschiedlicher Textsorten
  - Fragestellung gemäß [Hovy 2004: 594]:  
"Ask experts to underline and extract the most interesting or informative fragments of the text. Measure recall and precision of the system's summary against the human's extract ..."
  - jeweils 7 Texte mit/ohne Suchterme (Retrieval- versus Summary-Modus)
- Zur Menge der relevanten Sätze gehören alle, die von mehr als 50% der Bewerter extrahiert wurden.

# EXCERPT

---

## ■ Performance eigene Texte (Retrieval-Modus):

Text (Sätze)	BL-Initial	EXCERPT
1 (4/15)	0.50	1.00
2 (6/24)	0.00	0.50
3 (10/26)	0.50	0.90
4 (5/30)	0.20	1.00
5 (9/31)	0.20	0.78
6 (15/36)	0.47	0.73
7 (5/68)	0.20	0.40
∅ (7.7/33)	0.30	0.76

---

# EXCERPT

---

## ■ Ausblick:

### ● Weitere Performance-Tests:

- Quervergleich zu anderen Within-Document-Retrieval-Systemen und Summarizern
- Mehr und evtl. längere Texte mit Suchanfragen
- New-York-Times-Texte (Online-Summaries)

### ● Weitere Parameter:

- Überschriftenterme
- ?

---

Vielen Dank für Ihre  
Aufmerksamkeit!

# Literatur

---

- [Banko & al. 1999] Banko, M. & Mittal, V. & Kantrowitz, M. & Goldstein, J. (1999): Generating Extraction-Based Summaries from Hand-Written Summaries by Aligning Text-Spans. *Proceedings of the Pacific Association for Computational Linguistics PACLING*.
- [Edmundson 1969] Edmundson, H. P. (1969): New Methods in Automatic Ex-tracting. *Journal of the American Society for Information Science*, 16(2), S. 264–285.
- [Goldstein & al. 1999] Goldstein, J. & Kantrowitz, M. & Mittal, V. & Carbonell, J. (1999): Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of SIGIR'99*, S. 121–128.
- [Harper et al. 2004] Harper, D. J. & Koychev, I. & Sun, Y. & Pirie, I. (2004): Within-Document-Retrieval: A user-Centred Evaluation of Relevance Profiling. *Information Retrieval*, 7, S. 265-290.

# Literatur

---

- [Mani & Maybury 1999] Mani, I. & Maybury, M. T. (1999; Hrsg.): *Advances in Automatic Text Summarization*. Cambridge & London: MIT Press.
- [Mani 2001] Mani, I. (2001): *Automatic Summarization*. Amsterdam & Philadelphia: Benjamins.
- [Mittal & al. 1999] Mittal, V. & Kantrowitz, M. & Goldstein, J. & Carbonell, J. (1999): Selecting Text Spans for Document Summaries: Heuristics and Metrics. *Proceedings of AAAI-99*, S. 467–473.
- [Salton & al. 1993] Salton, G. & Allan, J. & Buckley, C. (1993): Approaches to Passage Retrieval in Full Text Information Systems. *Proceedings of ACM-SIGIR'93*, S. 49–58.

# Diverses

---

## ■ Passagen-Retrieval:

- Ranking von Dokumenten nach der jeweils relevantesten Passage statt nach der Relevanz des Gesamtdokuments (als Antwort auf eine Suchanfrage in der Dokumentenkollektion);
- Ranking von mehr oder weniger relevanten Passagen innerhalb eines Dokuments, das nicht aus einer Dokumentenkollektion stammen muss ('Within-Document-Retrieval');
- Retrieval der relevantesten Stelle aus allen Treffer-Dokumenten, die vermutlich die Antwort auf eine konkrete Faktenfrage enthält ('Question-Answering-Systeme').

# Diverses

---

- 'Expert in Computational Evaluation and Retrieval of Passages of Text'