UR

Informationslinguistik

Textanalyse II: Inhaltliche Analyse (Skript 2014)

Informationswissenschaft
Universität Regensburg
Jürgen Reischer

Einführung

- Auf Basis grundlegender Analyseschritte – *Tokenisierung*, *Normalisierung*, *Mehrwortausdruck-Erkennung*, *Eigennamen-Erkennung*, *Satzerkennung*, *POS-Tagging*, *N-Gramm-Ermittlung* – lassen sich weitergehende Textanalysen durchführen:
  - *Indexierung*: Ermittlung der relevantesten Inhaltsausdrücke aus einem Textdokument (s. unten);
  - *Zusammenfassung*: Ermittlung der relevantesten Information aus einem Textdokument (s. unten);
  - *Information-Retrieval*: Ermittlung der relevantesten Dokumente aus einem Dokumentenbestand.

Einführung

- Durch Indexierung und Zusammenfassung wird jeweils eine *Repräsentation* des Textes erzeugt:
  - Die Repräsentation ist dabei eine verkürzte und präzise Wiedergabe der wesentlichen Inhalte des Textes;
  - Indexierung und Zusammenfassung unterscheiden sich in der Wahl der Ausdrucksmittel:
    - *Indexierung*: isolierte Wörter und Phrasen aus dem Text und/oder vom Indexierer;
    - *Zusammenfassung*: neu konstruierter Text mit vollständigen und zusammenhängenden Sätzen aus dem Text und/oder vom Zusammenfasser.

Einführung

- Die Textrepräsentation muss verschiedene Aufgaben bzw. Bedingungen erfüllen:
  - Sie muss das Thema des Textes (engl. 'aboutness') unzweideutig wiedergeben:
    - Inhaltliche Mehrdeutigkeiten sind aufzulösen und formale Varianten zu normalisieren;
    - der 'topikale Fokus' des Textes ist zu ermitteln, um den Text von anderen Texten gleicher Thematik möglichst gut abgrenzen zu können.
  - Sie ist eine komprimierte/kondensierte Darstellung der Textinhalte.

Indexierung — Grundlagen

- Indexierung bezeichnet den Vorgang der Inhaltserschließung von Textdokumenten zu deren Verwaltung und Wiederauffindung:
  - Textdokumenten werden dabei Indexausdrücke (Indexterme) zugeordnet:
    - Dokumente werden in einer Dokumentenkollektion gespeichert und sollen bei Bedarf anhand bestimmter Suchbegriffe/Stichwörter wieder abrufbar sein;
    - die Erschließung findet dabei zumeist automatisiert statt, da die Vielzahl an Dokumenten die manuelle Auswertung nicht mehr erlaubt (z. B. im WWW).

Indexierung — Grundlagen

- Neben dem Dokument selbst werden oftmals auch zugehörige Texte wie die Zusammenfassung (das 'Abstract'), der Titel oder Kapitel-Überschriften (bzw. das Inhaltsverzeichnis) mitindexiert:
  - Gerade dort sind die Inhalte des zu indexierenden Dokuments besonders prägnant formuliert, da hier nur die wirklich relevanten Ausdrücke auftreten;
  - oftmals werden deshalb nur diese Texte stellvertretend für das Dokument als Ganzes zur Indexierung verwendet:
    - schnellere und platzsparendere Indexierung;
    - höhere Genauigkeit der Indexierung (weniger inhaltlicher 'Schmutz').

Indexierung — Thesaurus

- Neben den Originalbegriffen (Stichwörtern) aus dem Textdokument und/oder seinen Stellvertretern (Zusammenfassung) können zusätzlich/ersatzweise Ausdrücke aus einem allgemeinen *Thesaurus* verwendet werden:
  - Ein Thesaurus ist eine Art Schlagwörter-Lexikon mit einem *kontrollierten Vokabular*:
    - Die darin enthaltenen Ausdrücke werden inhaltlich eindeutig definiert und
    - mit anderen Ausdrücken über semantische Relationen in Beziehung gesetzt (z. B. via Synonyme, Hyp[er]onyme usw.);

Indexierung — Thesaurus

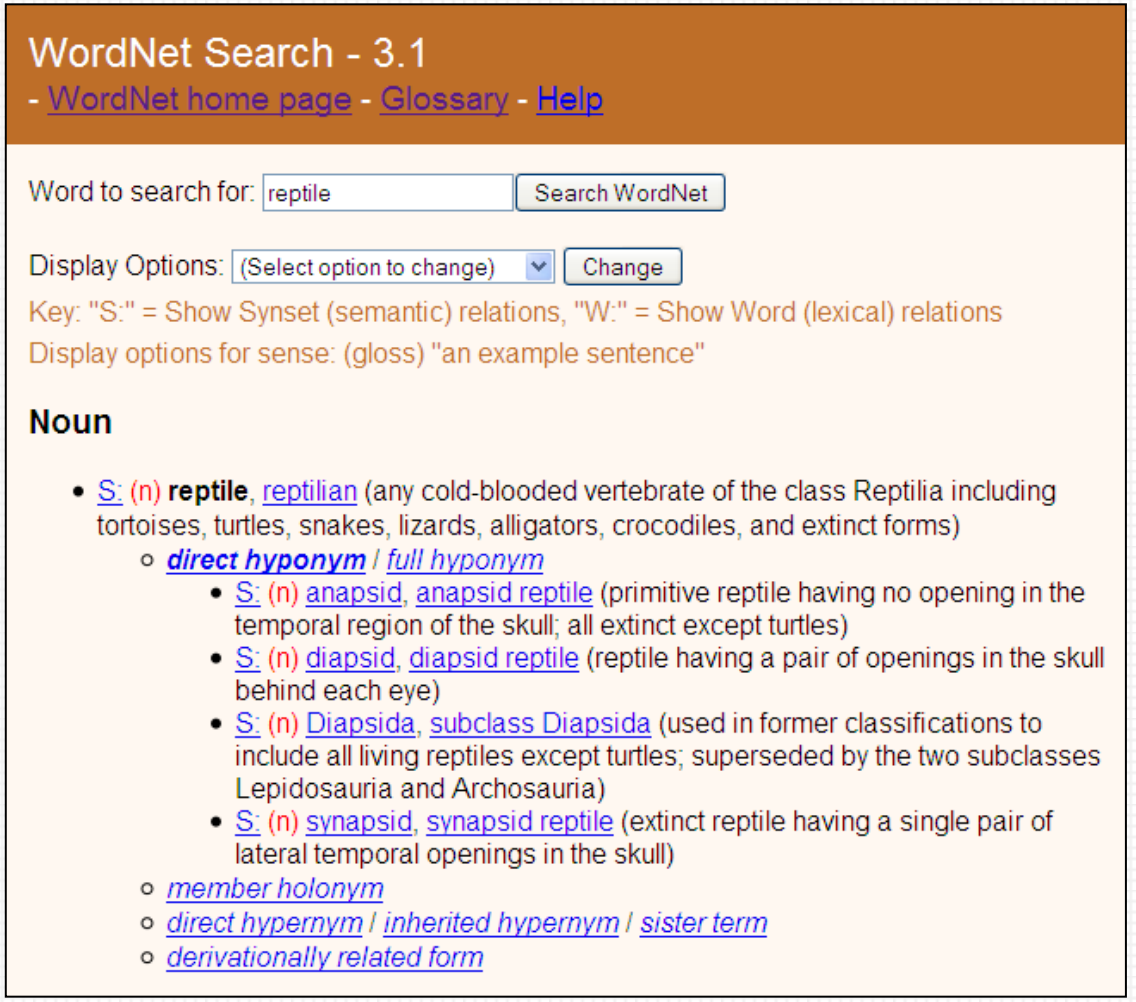
- da ein relevantes Konzept auf verschiedene Weisen sprachlich ausgedrückt werden kann, werden Index-ausdrücke durch inhaltlich verknüpfte Thesaurus-Ausdrücke ('Terme') erweitert:
  - bei der Indexierung des Dokuments bzw. seiner Stellvertreter selbst;
  - bei der Suche eines Anwenders mittels Suchbegriffen/Stich-wörtern.

  Dadurch kann sichergestellt werden, dass verschiedene Terminologien für das selbe zu bezeichnende Phänomen erfasst sind (z. B. "humour" vs. "humor", "holiday" vs. "vacation", "website" vs. "internet site" usw.).

Indexierung — Thesaurus

Beispiel WordNet-Thesaurus (18.12.13):

- Darstellung des Konzepts 'Reptil' mit den zwei englischen synonymen Ausdrücken "reptile" & "reptilian";
- dazu die vier Unterbegriffe unter 'direct hyponym' und weitere verknüpfte Begriffe (nicht geöffnet).



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: reptile [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **reptile**, reptilian (any cold-blooded vertebrate of the class Reptilia including tortoises, turtles, snakes, lizards, alligators, crocodiles, and extinct forms)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) anapsid, anapsid reptile (primitive reptile having no opening in the temporal region of the skull; all extinct except turtles)
    - S: (n) diapsid, diapsid reptile (reptile having a pair of openings in the skull behind each eye)
    - S: (n) Diapsida, subclass Diapsida (used in former classifications to include all living reptiles except turtles; superseded by the two subclasses Lepidosauria and Archosauria)
    - S: (n) synapsid, synapsid reptile (extinct reptile having a single pair of lateral temporal openings in the skull)
  - *member holonym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

Indexierung — Thesaurus

- In einem Thesaurus werden gleichbedeutende Ausdrücke (Synonyme) zu einem *Konzept* (Synonym-Menge) zusammengefasst:
  - Einer der Ausdrücke wird dabei als *Deskriptor* bestimmt, der als *bevorzugte* und *eindeutige Standardbezeichnung* verwendet wird;
  - andere ähnlich geschriebene oder gleichbedeutende Ausdrücke werden auf die Standardbezeichnung zurückgeführt, um eine einheitliche Dokument-repräsentation zu erreichen (Beispiele):

Indexierung – Thesaurus

- Formale Zurückführung (z. B. Schreibvarianten):
  - Abkürzung vs. Vollform ("GB" ⇨ "Great Britain");
  - Getrennt- vs. Zusammenschreibung ("web site" ⇨ "website");
  - Amerikanisches Englisch vs. Britisches Englisch ("cozy" ⇨ "cosy").
- Inhaltliche Zurückführung (z. B. Bedeutungsnuancen):
  - Konnotation vs. Denotation ("steed"/"nag" ⇨ "horse");
  - Wortbildung vs. Einfachausdruck ("morning star" ⇨ "Venus");
  - Fachterminus vs. Alltagsausdruck ("Equus caballus" ⇨ "horse");
  - Hochsprache vs. Standardsprache ("omnipotent" ⇨ "almighty").
- Dadurch wird 'terminologische Kontrolle' ausgeübt, so dass gleiche Dinge/Phänomene im Dokument mit denselben sprachlichen Mitteln ausgedrückt werden.

Indexierung — Informationslinguistik

- Bevor ein Ausdruck aus dem Text in den Index aufgenommen wird, durchläuft er verschiedene Phasen der informationslinguistischen Verarbeitung:
  - Vereinigung mit anderen Wörtern zu möglichen Mehrwortausdrücken, sofern das betrachtete Wort Teil eines solchen Mehrwortausdrucks ist (Phrasen sind informativer als die einzelnen Wörter);
  - Kategorienermittlung zum Ausschluss von Funktionswörtern via POS-Tagging oder Funktionswörter-Liste;
  - Normalisierung im Sinne von Deflexion und Dekomposition zur Ermittlung der Grundform/Zitierform, die im Index eingetragen werden soll (vgl. Buchindex).

Indexierung – Informationslinguistik

- Beispiel einer Normalisierung auf Grundformen:

Nicht-linguistische Grundform	Im Text auftretende Wortformen/Vollformen	Linguistische Grundform	Linguistische Stammform
"comput"	"compute(s/d/ing)"	"compute"	"compute"
	"computer(s)"		
	"computable(ly)"		
	"computation(s)"	"comput-ation"	
	"computational(ly)"		
"computeriz"	"computerize(s/d/ing)"	"computer-ize"	
	"computerizable(ly)"		
	"computerization(s)"	"computer-ization"	
	"computerizational(ly)"		

Indexierung — Informationslinguistik

- Grundsätzlich werden nur inhaltstragende Ausdrücke zur Indexierung verwendet, da nur sie das Thema eines Textes konstituieren können:
  - *Inhaltswörter*: Nomen, Verben, Adjektive, Adverben;
  - *Funktionswörter*: Adpositionen, Pronomen, Determinatoren, Junktionen, Partikeln.

  Der Ausschluss von Funktionswörtern erfolgt z. B. durch eine Negativliste, in der alle Funktionswörter einer Sprache aufgelistet sind (ca. 500). Zudem kann eine Negativliste für irrelevante oder häufige Inhaltswörter verwendet werden, die in allen Texten auftreten.

Indexierung — Gewichtungsfaktoren

- Ausdrücke können zudem nach ihrer *Diskriminanz* (Unterscheidungskraft) gewichtet werden:
  - Jeder Ausdruck leistet einen individuellen Beitrag zur inhaltlichen Erschließung und Wiederauffindbarkeit eines Dokuments:
    - Je diskriminativer (unterscheidungsfähiger) ein Ausdruck im Hinblick auf die inhaltliche Charakterisierung eines Dokuments ist, desto wichtiger ist er für dessen Index;
    - zu allgemeine und/oder zu häufig überall auftretende Ausdrücke charakterisieren den jeweils spezifischen Inhalt eines bestimmten Dokuments nur unzureichend.

Indexierung — Gewichtungsfaktoren

- Zwei Faktoren bestimmen die Diskriminanz eines Ausdrucks (engl. 'term'):
  - Häufigkeit des Auftretens des Ausdrucks im betrachteten Dokument (Termfrequenz *tf*):
    - Je häufiger der Ausdruck im Dokument auftritt, desto besser charakterisiert er dessen Inhalt und desto wichtiger ist er;
    - die häufige Wiederholung des Ausdrucks bedeutet, dass über das zugrundeliegende Konzept mehrfach an verschiedenen Textstellen gesprochen wurde ('roter Faden');
    - nur *ein*mal in einem Textdokument auftretende Ausdrücke sind für den Inhalt oder das Thema des Textes offenbar nicht relevant genug, um ausführlicher behandelt zu werden (oder es handelt sich um Orthografie- oder Tippfehler).

Indexierung — Gewichtungsfaktoren

- Häufigkeit des Auftretens des Ausdrucks in einem Korpus bzw. einer Dokumentenkollektion, zu der das betrachtete Dokument gehört (inverse Dokumentfrequenz *idf*):
  - Je häufiger der Ausdruck in der Dokumentenkollektion sonst noch vorkommt, desto weniger charakterisiert er inhaltlich speziell das betrachtete Dokument;
  - das häufige Auftreten eines Ausdrucks auch in anderen Dokumenten legt nahe, dass er in vielen anderen inhaltlich-thematischen Kontexten auftritt und daher nicht spezifisch für das betrachtete Dokument ist;
  - damit handelt es sich aber gerade um einen Ausdruck, der das betrachtete Dokument inhaltlich schlecht erschließt, da er zugleich auch noch charakteristisch für die Inhalte anderer Dokumente ist.

Indexierung — Gewichtungsfaktoren

Die beiden Faktoren tf und idf zusammen können nun gemeinsam als Gewichtungsfaktor für einen Ausdruck t herangezogen werden:

- tf ist definiert als die absolute Frequenz des betrachteten Ausdrucks (Anzahl Term-Vorkommnisse im Dokument) FQ_t in Relation zur Gesamtzahl Ausdrücke im Dokument FQ_d, d. h. es gilt $tf = FQ_t / FQ_d$:
  - *Je größer FQ_t wird, desto größer wird tf*: Je häufiger der Ausdruck im Dokument auftritt, desto besser charakterisiert er den Inhalt.
  - *Je größer FQ_d wird, desto kleiner wird tf*: Je mehr Ausdrücke im Dokument überhaupt vorhanden sind, desto geringer ist der Anteil von t an allen Dokumentausdrücken, d. h. desto weniger ist er charakteristisch für den Inhalt (da es noch andere Terme gibt).

Indexierung — Gewichtungsfaktoren

- idf ist definiert als der Logarithmus der Gesamtzahl an Dokumenten FQ_n in der Kollektion in Relation zur Gesamtzahl Dokumente FQ_k, in denen der betrachtete Term auftritt, d. h. es gilt idf = log (FQ_n / FQ_k):
  - *Je größer FQ_n wird, desto größer wird idf*: Je mehr Dokumente im Dokumentenbestand vorhanden sind, desto diskriminativer wird der betrachtete Term für diejenigen Dokumente, in denen er tatsächlich auftritt.
  - *Je größer FQ_k wird, desto kleiner wird idf*: In je mehr Dokumenten der betrachtete Term vorkommt, desto weniger diskriminativ ist er für die Dokumente, in denen er auftritt.

  Der Logarithmus des Bruches FQ_n / FQ_k dient nur dazu, die Größenordnung an den tf-Wert anzugleichen (s. u.).

Indexierung — Gewichtungsfaktoren

Die beiden Faktoren tf und idf werden in der Regel zu einem gemeinsamen Gewichtungsfaktor w = tf·idf multipliziert:

- Für einen betrachteten Ausdruck t hängt dessen Gewichtung w insgesamt von vier frequenzbasierten Größen ab:
  - der Anzahl Vorkommnisse FQ_t des Terms t im Dokument;
  - der Gesamtzahl aller Terme FQ_d im betrachteten Dokument;
  - der Gesamtzahl aller Dokumente FQ_n in der Kollektion;
  - der Gesamtzahl aller Dokumente FQ_k, in denen t auftritt.
- Das Gewichtungsmaß tf·idf hat den Vorteil, dass generell häufig auftretende Funktions- und Inhaltswörter automatisch heruntergewichtet werden.

Indexierung — Herausforderungen

- Herausforderungen der automatischen Indexierung:
  - Auflösung von Pronomen (als Funktionswörter) durch Ersetzung ihrer Bezugsausdrücke (Inhaltswörter, meist Nomen);
  - Auflösung von Ambiguitäten bei mehrdeutigen Ausdrücken zur Ermittlung des eigentlichen Deskriptors, der grundsätzlich eindeutig sein muss;
  - Umgang mit Ausdrücken aus anderen Sprachen, z. B. aufgrund eingebetteter Zitate (sind Dokumente mit gesuchten Ausdrücken in fremdsprachlichen Zitaten relevant?).

Übung – Vertiefung

- Indexieren Sie nachfolgenden Auszug aus dem Beispieltext. Überlegen und begründen Sie:
  - Worum geht es in dem Text, was ist das Thema?
  - Welche Ausdrücke aus dem Text spiegeln das Thema wider, welche Ausdrücke sollten eventuell hinzugefügt werden?
  - Welche Ausdrücke bzw. Kombinationen von Ausdrücken sind prägnant speziell für diesen Text, nicht aber für andere Texte ähnlicher Thematik?
  - Welche formalen Varianten der gewählten Indexausdrücke sind zu verwenden (Normalisierung)?
  - Welche Ausdrücke sind gegenüber anderen vorzuziehen (Deskriptoren)?
  - Welche Ausdrücke werden grundsätzlich nicht verwendet?

Übung – Vertiefung

Coldblooded Does Not Mean Stupid

Humans have no exclusive claim on intelligence. Across the animal kingdom, all sorts of creatures have performed impressive intellectual feats. A bonobo named Kanzi uses an array of symbols to communicate with humans. Chaser the border collie knows the English words for more than 1,000 objects. Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

And reptiles? Well, at least they have their looks.

In the plethora of research over the past few decades on the cognitive capabilities of various species, lizards, turtles and snakes have been left in the back of the class. Few scientists bothered to peer into the reptile mind, and those who did were largely unimpressed.

Übung – Vertiefung

Indexterme:

Zusammenfassung — Grundlagen

- Die maschinelle (automatische) Zusammenfassung von Text(dokumenten) bezeichnet den Prozess der Analyse und Synthese von Text, um
  - die wesentlichste Information aus einer gegebenen Menge an Text(en) zu ermitteln und
  - die ausgewählte Information zu einem informationell kondensierten, neuen Text zusammenzuführen.

  } Transformation von Information

Im Englischen wird die Zusammenfassung als 'Summary' bzw. 'Summarizing' bezeichnet.

Zusammenfassung — Grundlagen

- Die Zusammenfassung muss dabei bestimmte Bedingungen erfüllen:
  - Form:
    - Die Zusammenfassung ist kürzer als der Originaltext (gemessen in Zeichen, Wörtern, Sätzen etc.);
    - sie muss trotz Verkürzung jedoch immer noch lesbar sein, um dann auch inhaltlich verständlich sein zu können;
    - sie sollte denselben Ausdrucksstil hinsichtlich Struktur-aufbau und Terminologie verwenden;
    - sie sollte nicht den Titel des Originaltextes enthalten (da die Zusammenfassung oft zum Original selbst gehört).

Zusammenfassung — Grundlagen

- *Inhalt*:
  - Der Informationsgehalt des Textkondensats darf die Informationsmenge des Originaltextes nicht übersteigen (d. h. keine Erzeugung zusätzlicher Information, selbst wenn sie im Sinne des Originals wäre);
  - die Information im Summary sollte prägnant sein und die Inhalte des Originals auf den Punkt bringen (d. h. Entfernung irrelevanter und redundanter Information aus dem Original);
  - die Zusammenfassung sollte die Meinung/Haltung des Originalautors wiedergeben, d. h. keine Verschiebung der Akzente des Originals (z. B. Zitate, Fachtermini usw.);
  - die Information im Summary sollte mindestens ebenso verständlich bzw. verständlicher sein als im Original.

Zusammenfassung — Grundlagen

- Einsatzzweck von Zusammenfassungen:
  - Ausblick auf bzw. Überblick über Textinhalte:
    - Unterstützung bei der Beurteilung der inhaltlichen bzw. thematischen Relevanz des Originaldokuments;
    - Stellvertretung des Originaltextes zur Lesezeit-Ersparnis für Rezipienten;
    - Erleichterung/Verbesserung der maschinellen Analyse im Hinblick auf Indexierung.
  - Rückblick auf Textinhalte:
    - Verbesserung der Behaltensleistung von Textinhalten;
    - Reformulierung schwieriger Sachverhalte.

Zusammenfassung — Wichtigkeit/Relevanz

- Welche Information des Originaltextes als 'wichtig' oder 'relevant' gilt, ist je nach Anwendung/Anwender der Zusammenfassung festzulegen:
  - Wichtigkeit/Relevanz bezüglich eines bestimmten Informationsbedürfnisses eines Nutzers in einer Nutzungssituation (*subjektives Interessenprofil*):
    - Angabe von Themabegriffen, die spezielles Informationsbedürfnis ausdrücken: *Wofür interessiert sich der Nutzer*?
    - Angabe von Parametern, die Nutzungssituation charakterisieren (z. B. Länge und Art der Zusammenfassung): *Wofür und wie setzt der Nutzer die Information ein*?

Zusammenfassung — Wichtigkeit/Relevanz

- Wichtigkeit/Relevanz im Hinblick auf allgemeine (potenzielle, antizipierte) Nutzer und Nutzungs-szenarien (*objektives Interessenprofil*):
  - Thematischer Schwerpunkt des Textes als allgemeines Informationsbedürfnis (ohne Fokus auf Teilthemen oder spezielle Begriffe): *Worüber gibt der Text grundsätzlich Auskunft? Was wollte der Autor sagen?*
  - Neutralität im Hinblick auf Nutzungsszenarien (ohne Eingriff in bestimmte Merkmale der Zusammenfassung wie Länge): *Wie könnte ein Nutzer in einer nicht näher umrissenen Anwendungssituation die Zusammenfassung zu welchem Zweck nutzen wollen?*

Zusammenfassung — Klassifikation

- Beispiele für Zusammenfassungen aller Art:
  - Uneigenständige Zusammenfassungen von Textteilen innerhalb eines Dokuments:
    - *Inhaltsverzeichnis*: Prospektiver Überblick über die Themen eines Dokuments (bestehend aus den Überschriften einzelner Abschnitte);
    - *Überschriften/Titel*: Prospektive Zusammenfassung des Themas eines Abschnitts oder des ganzen Dokuments;
    - *Indexe*: Retrospektiver Abriss der wichtigsten Themabegriffe eines Dokuments (zweckgebunden zum Auffinden relevanter Textstellen);

Zusammenfassung — Klassifikation

- *Glossar*: pro-/retrospektive Zusammenfassung der Verwendungsweisen thematisch wichtiger Begriffe (zweckgebunden zur Erläuterung wiederkehrender Themabegriffe);
- *Literaturangaben*: retrospektiver Überblick über die behandelten Begriffe/Themen (zweckgebunden zur Verweisung auf Teilthemen);
- *Abstract/Extract*: prospektive Zusammenfassung der Inhalte eines Dokuments (zweckgebunden z. B. zur Ermittlung der Relevanz eines Dokuments);
- *Abschnittszusammenfassungen*: retrospektiver Rückblick auf die zentralen Aussagen eines Abschnitts.

Zusammenfassung — Klassifikation

- Eigenständige Zusammenfassungen von Dokumenten (nicht integraler Bestandteil eines Textes):
  - *Referat*: Wiedergabe der wesentlichen Inhalte eines Dokuments oder einer Dokumentenmenge in eigenen Worten, eventuell auch kritisch kommentierend oder hinterfragend (Referat: Verweisung auf Original[e]);
  - *Rezension*: kritisch-wertende Wiedergabe eines Dokuments zum Zwecke der (Nicht-)Empfehlung eines Werkes;
  - *Lehrbuch*: einführender Überblick über einen Themenbereich, den aktuellen Status Quo wiedergebend;
  - *Protokoll*: verdichtende Mitschrift einer vor allem mündlich ausgetragenen Konversation.

Zusammenfassung — Klassifikation

- Charakterisierungsdimensionen für verschiedene Arten von Summarys:
  - *indikativ vs. informativ*: den Inhalt andeutend vs. die wesentliche Information wiedergebend:
    - indikative Summarys sollen helfen einzuschätzen, ob das Dokument relevant ist und die Rezeption des Originals lohnt (die Inhalte des Originals werden kurz vorgestellt);
    - informative Summarys sollen das Originaldokument idealerweise ersetzen, so dass dessen Rezeption gerade nicht mehr notwendig ist (d. h. die wesentlichen Inhalte werden gleich wiedergegeben).

Zusammenfassung — Klassifikation

- *abstraktiv vs. extraktiv*: den Textinhalt in eigenen Worten reformulierend vs. wörtlich exzerpierend:
  - abstraktive (derivative) Summarys:
    - Wiedergabe der Inhalte aus der Perspektive des Abstrakt-Verfassers (bzw. Programmierers des Abstracting-Systems);
    - *sinngemäße* (konzeptuelle) Umformulierung der Originalinhalte als eigene geistige Leistung des Abstract-Autors.
  - extraktive Summarys:
    - Wiedergabe der Inhalte aus der Perspektive des Originaltextes bzw. -autors (bzw. Programmierers des Extracting-Systems);
    - *wörtliche* auszugsweise Übernahme von Originalformulierungen und Zusammenstellung zu neuem Text (z. B. Auswahl einzelner Sätze des Originaldokuments, die neu verbunden werden).

Zusammenfassung — Klassifikation

- *neutral vs. evaluativ*: den Inhalt wertungsfrei wiedergebend vs. kritisch-kommentierend darstellend:
  - neutrale Summarys sollen die Textinhalte sachlich-objektiv wiedergeben (gerade bei strittigen Themen wie politischen oder wissenschaftlichen Denkrichtungen);
  - evaluative Summarys dienen der subjektiv-wertenden Kurzdarstellung von Textinhalten gemäß der Meinung des Summary-Verfassers:
    - verschiedene Verfasser können entsprechend zu unterschiedlichen Darstellungen und Wertungen gelangen;
    - neutrale Summarys hingegen sollten von allen Verfassern idealerweise ähnlich konzipiert sein.

Zusammenfassung — Klassifikation

- *generisch vs. spezifisch*: nutzer-unspezifisch vs. nutzerzentriert zusammenfassend (s. oben):
  - Generische Summarys stellen unabhängig von speziellen Themabegriffen und/oder Anwendungsszenarien die wesentliche Information aus einem Text bereit;
  - spezifische Summarys orientieren sich an den Nutzer-bedürfnissen bzw. der Nutzungssituation.
- *individuell vs. aggregativ*: ein vs. mehrere Dokumente nicht-redundant zusammenfassend:
  - individuelles Summary eines einzelnen Dokuments;
  - aggregatives Summary von Dokumenten zu einem Thema.

Zusammenfassung — Beispiele

- Beispiele für verschiedene Zusammenfassungen anhand "Reptile"-Text (s. Anhang 'Textanalyse I'):
  - generisch-informatives Extract-Summary (beliebige 7 aufeinanderfolgende Sätze):

> The research could not only redeem reptiles but also shed new light on cognitive evolution. Because reptiles, birds and mammals diverged so long ago, with a common ancestor that lived 280 million years ago, the emerging data suggest that certain sophisticated mental skills may be more ancient than had been assumed - or so adaptive that they evolved multiple times.

Zusammenfassung — Beispiele

For evidence of reptilian intelligence, one need look no further than the maze, a time-honored laboratory test. Anna Wilkinson, a comparative psychologist at the University of Lincoln in England, tested a female red-footed tortoise named Moses in the radial arm maze, which has eight spokes radiating out from a central platform. Moses' task was to "solve" the maze as efficiently as possible: to snatch a piece of strawberry from the end of each arm without returning to one she had already visited. "That requires quite a memory load because you have to remember where you have been," Dr. Wilkinson said. Moses managed admirably, performing significantly better than if she had been choosing arms at random.

Zusammenfassung — Beispiele

- spezifisch-informatives Extract-Summary mit Thema-begriffen "reptile" & "cognition" (5 selektive Sätze):

> Few scientists bothered to peer into the reptile mind, and those who did were largely unimpressed. But now that is beginning to change, thanks to a growing interest in "coldblooded cognition" and recent studies revealing that reptile brains are not as primitive as we imagined. The research could not only redeem reptiles but also shed new light on cognitive evolution. Scientists say that many early studies of reptile cognition, conducted in the 1950s and '60s, had critical design flaws. The field of reptile cognition is in its infancy, but it already suggests that "intelligence" may be more widely distributed through the animal kingdom than had been imagined.

Zusammenfassung — Beispiele

- evaluativ-indikatives Abstract-Summary:

> The article shows that science has neglected some basic facts about reptiles concerning their intelligence. New insights and results from scientific investigations are presented that contrast reptilian with mammalian intelligent behavior. Experiments are described which reveal more flexible and context-appropriate reptile behavior than previously assumed.
> The article gives a short but interesting overview of new scientific research in the field of reptilian intelligence, and sheds light on obviously misguided scientific processes in the past.

Zusammenfassung — Parameter

- Durch Stellgrößen (Parameter) lässt sich ein Summary in verschiedener Hinsicht anpassen (z. B. bezüglich Nutzerpräferenzen):
  - Länge des Summarys in sprachlichen Einheiten (Zeichen, Wörter, Sätze, Absätze) bezogen auf die Länge des Originals:
    - explizite Längenangabe durch Nutzer, der absolute Anzahl N oder relative Menge P% an Einheiten wünscht;
    - implizite Längenbestimmung durch System, das aufgrund seiner Analysen selbst über den Umfang des Summarys bestimmt.

Zusammenfassung — Parameter

- Kleinste sprachliche Einheit zur Erstellung eines Summarys bestimmter Länge:
  - Je kleiner die gewählte Einheit, desto genauer lässt sich der gewünschte Umfang des Summarys erreichen (z. B. 500 Zeichen vs. 100 Wörter vs. 5 Sätze vs. 1 Absatz);
  - Sätze/Absätze können in ihrer Länge stark variieren, so dass sich kürzere Einheiten aus weniger Wörtern besser zusammenstellen lassen;
  - die minimale und/oder maximale Länge eines Summarys lässt sich in jedem Fall nur annähernd erreichen, da sonst Sätze oder Absätze gekürzt werden müssten (Entfernung von Teilsätzen erfordert gewissen Analyseaufwand).

Zusammenfassung — Parameter

- Art der Zusammenstellung von Einheiten im Hinblick auf die Abfolge im Summary vs. Originaltext (bei extraktivem Summary):
  - Beibehaltung der Reihenfolge von Einheiten entsprechend dem Originaltext, um inhaltliche Ablauflogik zu wahren;
  - Auswahl kontinuierlicher (direkt aufeinanderfolgender) vs. diskontinuierlicher (verstreuter) Einheiten:
    - kontinuierlich: bessere Verständlichkeit, jedoch evtl. geringerer Informationswert (da evtl. auch irrelevante Einheiten darunter);
    - diskontinuierlich: schlechtere Verständlichkeit, jedoch evtl. höherer Informationswert (da nur die wirklich wichtigen Einheiten übernommen werden).

Zusammenfassung — Kriterien

- Durch verschiedene Mechanismen wird versucht, die formale Länge des Originals unter möglichst weitgehender Erhaltung des Inhalts zu reduzieren:

***Ersetzen* von Einheiten**	***Entfernen* von Einheiten**
formal: Austauschen längerer Einheiten durch kürzere (Substitution)	*formal*: Auslassen von Einschüben in Parenthesen (Reduktion)
inhaltlich: Generalisieren spezifischer/detaillierter Information zu allgemeiner Information (Abstraktion)	*inhaltlich*: Auslassen irrelevanter/redundanter Information und Beibehaltung wesentlicher Information (Extraktion)

Zusammenfassung – Kriterien

- Formale Ersetzung von Einheiten:
  - Lexikalische Ersetzungen (Synonyme):
    - "Volkswagen" ⇨ "VW";
    - "Albert Einstein ⇨ "Einstein";
    - "search for extraterrestrial intelligence" ⇨ "SETI";
    - "test candidate"/"examinee" ⇨ "testee".
  - Grammatikalische Ersetzungen (Umschreibungen):
    - "under normal circumstances" ⇨ "normally";
    - "with respect to" ⇨ "concerning";
    - "extraordinarily big" ⇨ "very big" / "huge".

Zusammenfassung – Kriterien

- Inhaltliche Ersetzung von Einheiten:
  - Verallgemeinerung (Oberbegriffe):
    - "blue Levis-Jeans" ⇨ "Jeans";
    - "apples, bulbs, and plums" ⇨ "fruits";
    - "get influenza" ⇨ "sicken".
  - Verdichtung (Sammelbegriffe):
    - "French and Italian Riviera" ⇨ "Riviera";
    - "peel and cut potatoes, then boil them in salted water" ⇨ "boil potatoes";
    - "rich and poor, old and young people" ⇨ "all people".

Zusammenfassung — Kriterien

- Formale Entfernung von Einheiten:
  - Aussagelose Einheiten (z. B. weniger als 4 Wörter):
    - "Who was Einstein?";
    - "Know Einstein!".
  - Eingeschobene Einheiten (Parenthese-Strukturen):
    - "Einstein, the well-known physicist, discovered …" ⇨ "Einstein discovered …";
    - "Einstein – one of the world's greatest physicists –, discovered …" ⇨ "Einstein discovered …";
    - "Einstein (pronounced ['ɪnstɪn]) discovered …" ⇨ "Einstein discovered …".

Zusammenfassung — Kriterien

- Untergeordnete Einheiten (Nebeninformation):
  - "Einstein discovered that energy ist equivalent to mass, *i. e. the famous formula* $E = m \cdot c^2$." ⇨ "Einstein discovered that energy ist equivalent to mass."
  - "Einstein discovered that energy ist equivalent to mass, *although a normal person is unable to understand the consequences of this pioneering insight*." ⇨ "Einstein discovered that energy ist equivalent to mass.";
  - "The name 'Einstein' is used a stem in word formation, *e. g. in derived complex words like 'Einsteinium' or 'Einsteinian'*." ⇨ "The name 'Einstein' is used (as a stem) in word formation.".

Übung – Vertiefung

- Kürzen Sie folgenden Text gemäß der oben angegebenen Methoden:

> Was ist eine Wasservergiftung?
>
> Wer in wenigen Stunden fünf oder mehr Liter Wasser trinkt, schwebt in Lebensgefahr. Durch die große Flüssigkeitsmenge, die aus dem Verdauungstrakt in den Blutkreislauf aufgenommen wird, gerät der Salz-Haushalt aus dem Gleichgewicht, und im Gehirn bilden sich Wasserablagerungen. Es entsteht ein Überdruck, zunächst spürt man leichte Kopfschmerzen. Orientierungslosigkeit setzt ein, und schließlich kommt es zu Atemnot, Nierenversagen, Bewusstlosigkeit oder schlimmstenfalls sogar zum Tod. Tipp der Mediziner: Zwei bis maximal drei Liter Wasser über den Tag verteilt sind völlig ausreichend.

tv14, Nr. 8/2014, S. 22

Zusammenfassung — Kriterien

- Inhaltliche Entfernung von Einheiten:
  - Für die inhaltliche Entfernung von Einheiten ist eine komplexere Inhaltsanalyse notwendig:
    - Zunächst ist die relevante vs. irrelevante Information zu ermitteln und Letztere zu eliminieren;
    - aus der verbleibenden relevanten Information muss die redundante (doppelte) Information eliminiert werden.
  - Die inhaltlichen Eliminationsprozesse beziehen sich meist auf ganze Sätze oder Absätze (weniger auf Teilsätze oder einzelne Wörter), wodurch auch in die Gesamtstruktur des Textes eingegriffen wird.

Zusammenfassung — Kriterien

- Entscheidend ist die Bewertung von Einheiten wie Sätzen oder Absätzen nach ihrer thematischen Relevanz:
  - Im Hinblick auf extraktives Zusammenfassen werden vor allem Texteinheiten wie Sätze und Absätze anhand bestimmter Kriterien auf (Ir-)Relevanz hin bewertet;
  - Absätze bestehen aus Sätzen, die wiederum aus Wörtern (bzw. Mehrwortausdrücken) bestehen:
    - Die Bewertung einzelner Wörter ermöglicht die Gesamtbewertung von Sätzen und Absätzen;
    - die Bewertung von Sätzen wiederum ermöglicht die Gesamt-bewertung von Absätzen.

Zusammenfassung — Kriterien

Kriterien zur Bewertung von (Mehrwort-)Ausdrücken:

<table>
<tr><th>Kriterium</th><th colspan="2">Beschreibung</th></tr>
<tr><td rowspan="2">Kategorie des Ausdrucks</td><td>Inhaltswörter:
Sie tragen inhaltliche Bedeutung und leisten so einen Beitrag zum Thema des Textes.

Beispiele:
Nomen, Verben (Vollverben), Adjektive, Adverben (alle Hauptwortarten).</td><td>Funktionswörter:
Sie tragen formale Bedeutung und liefern keinen Beitrag zum Textgehalt/-thema.

Beispiele:
Adpositionen, Junktionen, Partikeln, Pronomen, Artikel, Hilfsverben, Interjektionen.</td></tr>
<tr><td colspan="2">Sätze mit vielen Inhaltswörtern sind deshalb grundsätzlich relevanter als solche mit vielen Funktionswörtern.</td></tr>
</table>

Zusammenfassung – Kriterien

- Beispieltext mit vielen Funktionswörtern (rot):

> Well, at least they have their looks.
>
> So how did we miss this for so long?
>
> Not one tortoise figured this out on its own.

- Beispieltext mit vielen Inhaltswörtern (grün):

> Humans have no exclusive claim on intelligence. A bonobo named Kanzi uses an array of symbols to communicate with humans. Chaser the border collie knows the English words for more than 1,000 objects. Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

Zusammenfassung – Kriterien

Kriterium	Beschreibung	
Indikatorausdrücke	*Bonusausdrücke*: Ausdrücke, die wichtige/interessante Inhalte anzeigen. *Beispiele*: Zusammenfassungsausdrücke ("in summary", "overall"); Relevanzausdrücke ("thus", "it is important to note"); Steigerungsausdrücke ("more"/"most", "less"/"least")	*Malusausdrücke*: Ausdrücke, die unwichtige/uninteress. Inhalte anzeigen. *Beispiele*: Beiläufigkeitsausdrücke ("well", "by the way", "e.g.") Irrelevanzausdrücke ("needless to say that") Unschärfeausdrücke ("more or less", "several")
	Sätze mit vielen Bonus- und wenigen Maluswörtern sind daher relevanter als solche mit umgekehrtem Verhältnis.	

Zusammenfassung – Kriterien

* Beispieltext mit Malusausdrücken (rot):

> For instance, scientists commonly use "aversive stimuli," such as loud sounds and bright lights, to shape rodent behavior.
>
> Scientists may also have been asking reptiles to perform impossible tasks.

* Beispieltext mit Bonusausdrücken (grün):

> Chaser the border collie knows the English words for more than 1,000 objects.
>
> Navigational skills are important, but the research also hints at something even more impressive: behavioral flexibility, or the ability to alter one's behavior as external circumstances change.

Zusammenfassung — Kriterien

<table>
<tr><th>Kriterium</th><th colspan="2">Beschreibung</th></tr>
<tr><td rowspan="2">Topik-ausdrücke</td><td>Satzinitiale Ausdrücke:
Wörter am Anfang eines Satzes sind thematisch zentraler als solche am Ende.

Am Satzanfang steht meist das grammatische Subjekt und/oder thematische Topik, d. h. tendenziell das Thema des Satzes/Textes.</td><td>Absatzinitiale Ausdrücke:
Sätze am Anfang eines Absatzes leiten das Thema des Absatzes bzw. nachfolgenden Abschnitts ein.

Sie sind für das Verständnis des nachfolgenden Textes unerlässlich und damit per se zentral.</td></tr>
<tr><td colspan="2">Wörter am Anfang einer Texteinheit (Satz/Absatz/Text) sind wichtiger für das Thema eines Textes als solche, die am Ende stehen.</td></tr>
</table>

Zusammenfassung — Kriterien

* Beispieltext mit satzinitialen (Subjekt-)Ausdrücken (grün):

> Across the animal kingdom, all sorts of creatures have performed impressive intellectual feats. A bonobo named Kanzi uses an array of symbols to communicate with humans. Chaser the border collie knows the English words for more than 1,000 objects. Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

* Beispieltext mit absatzinitialen Ausdrücken (grün):

> Humans have no exclusive claim on intelligence.
>
> Because reptiles, birds and mammals diverged so long ago, …
>
> For evidence of reptilian intelligence, one need look no further …

Zusammenfassung — Kriterien

<table>
<tr><th>Kriterium</th><th colspan="2">Beschreibung</th></tr>
<tr><td rowspan="2">Infor-
mativ-
ausdrü-
cke 1</td><td>Informativwörter:
Wörter, die informations-haltig im Sinne von seman-tisch spezifisch und präzise sind, sind wichtiger als allge-meine und unscharfe/vage Ausdrücke.

Beispiele:
Hyponyme vs. Hyperonyme,
Komposita vs. Einzelwörter</td><td>Informativsätze:
Sätze, die Aussagen machen, sind wichtiger als Sätze, die Fragen oder Aufforderungen beinhalten (Fragen und Auf-forderungen geben selbst keine Information ab).

Beispiele:
Sätze mit "." vs. Sätze mit "!" oder "?" am Ende.</td></tr>
<tr><td colspan="2">Sätze/Absätze mit vielen Informativausdrücken sind wichtiger als Einheiten mit weniger.</td></tr>
</table>

Zusammenfassung – Kriterien

* Beispieltext mit Nicht-Informativausdrücken (rot):

> And reptiles?
>
> Things became even more interesting when Dr. Wilkinson hung …

* Beispieltext mit Informativausdrücken (grün):

> Coldblooded Does Not Mean Stupid
>
> Chaser the border collie knows the English words …
>
> Anna Wilkinson, a comparative psychologist at the University of Lincoln in England, tested a female red-footed tortoise named …
>
> Anole, a tropical lizard, …

Zusammenfassung — Kriterien

Kriterium	Beschreibung
Informativausdrücke 2	Sätze mit vielen zum ersten Mal im Text angeführten Wörtern sind informativer als solche, deren Wörter im Text wiederholt werden (und damit redundant sind): • unbekannte Wörter, die bezüglich Referenzlexikon völlig neu sind (z. B. Komposita); • bekannte Wörter, die im betrachteten Satz jedoch zum ersten Mal im Text verwendet werden. Sätze mit vielen Neuwörtern verweisen im Text inhaltlich nicht zurück und stehen daher am Anfang eines neuen thematischen Abschnitts. Solche Sätze sind vor allem für das Verständnis des nachfolgenden Textes wichtig.
	Sätze/Absätze mit vielen Informativausdrücken sind wichtiger als Einheiten mit weniger.

Zusammenfassung – Kriterien

- Beispieltext mit Informativausdrücken (grün):

> Coldblooded Does Not Mean Stupid
>
> Humans have no exclusive claim on intelligence. Across the animal kingdom, all sorts of creatures have performed impressive intellectual feats. A bonobo named Kanzi uses an array of symbols to communicate with humans. Chaser the border collie knows the English words for more than 1,000 objects. Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

- Beispieltext mit Nicht-Informativausdrücken (rot):

> Now that scientists have gotten better at designing experiments for reptiles, they are uncovering all kinds of surprising abilities.

Zusammenfassung — Kriterien

Kriterium	Beschreibung
Relationsausdrücke	Ausdrücke in lexikalischer (semantisch-thematischer) und grammatikalischer Relation zu anderen Ausdrücken sind wichtiger als formal/inhaltlich unverbundene Ausdrücke. Sätze/Absätze mit solchen Ausdrücken sind mit anderen Sätzen/Absätzen kohäsiv bzw. kohärent und daher zentral. *Beispiele für grammatikalische Relationen (Kohäsion)*: Proformen, Konnektoren, Determinatoren, Kollokationen *Beispiele für lexikalische Relationen (Kohärenz)*: Synonymie, Antonymie, Hyp(er)onymie, Holo-/Meronymie, Assoziationen (alle Relationen, die nicht unter die genannten sprachlich-begrifflichen fallen, z. B. ontologische Kookkurrenzen zwischen Objekten/Sachverhalten)

Zusammenfassung – Kriterien

- Beispieltext mit grammatikalischen Relationen (eigene Farbe pro formal-kohäsivem Phänomen):

> Moses managed admirably, performing significantly better than if she had been choosing arms at random. Further investigation revealed that she was not using smell to find the treats. Instead, she seemed to be using external landmarks to navigate, just as mammals do.

- grün: Referenzkette mit vierfachem (pro)nominalen Bezug auf Moses ("red-footed tortoise");
- blau: Elliptisch/pronominale Wiederaufnahme eines Vorgangs;
- rot: Anschlussausdrücke (Konnektoren);
- orange: Kollokationen (evtl. auch "reveal that", "just as");
- grau: Determinatoren (Wiederaufnahme eines bereits eingeführten Konzepts, hier das 'Festessen' Erdbeere).

Zusammenfassung — Kriterien

- Beispieltext mit lexikalischen Relationen (eigene Farbe pro inhaltlich-kohärentem Strang):

> Humans have no exclusive claim on intelligence. Across the animal kingdom, all sorts of creatures have performed impressive intellectual feats. A bonobo named Kanzi uses an array of symbols to communicate with humans. Chaser the border collie knows the English words for more than 1,000 objects. Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

- grün: Tierreich;
- rot: Sprache/Kommunikation;
- orange: Intellekt/Kognition/Perzeption;
- grau: Objekte/Artefakte.

Zusammenfassung — Kriterien

Kriterium	Beschreibung
Titel-ausdrücke	Ausdrücke, die in speziellen Abschnitten des Textes stehen, sind (un)wichtiger als andere. *Beispiele für spezielle Abschnitte*: positiv: Überschriften, Titel, Aufzählungen, Resümees negativ: Parenthese-Konstruktionen (Klammerungen)
Frequenz-ausdrücke	Ausdrücke, die häufig in einem Text auftreten, *ohne zugleich generell in Texten bzw. der Sprache häufig aufzutreten*, sind wichtiger als seltene Ausdrücke. *Beispiel für generell häufige Ausdrücke (unbrauchbar)*: die meisten Funktionswörter, spezielle Inhaltswörter wie "people", "time", "say", "get", "go", "make", "good", "so" etc.

Zusammenfassung — Kriterien

- Beispieltext mit speziellen Abschnitten (⊕ grün, ⊖ rot):

Coldblooded Does Not Mean Stupid

Crows make sophisticated tools, elephants recognize themselves in the mirror, and dolphins have a rudimentary number sense.

(Unable to reach the reward, some of the animals simply decided to nap.)

Other studies of reptiles have turned up similar results, challenging the popular theory that social learning evolved as a byproduct of – and a special adaptation for – group living. Instead, Dr. Wilkinson said, social learning may be merely an outgrowth of an animal's general ability to learn.

Zusammenfassung — Kriterien

- Beispieltext mit generell häufigen Ausdrücken (rot):

"That requires quite a memory load because you have to remember where you've been," Dr. Wilkinson said.

This flexibility, which allows animals to take advantage of new environments or food sources, has been well documented in birds and primates, and scientists are now beginning to believe that it exists in reptiles, too.

So how did we miss this for so long?

To get their snouts on the treat, the tortoises needed to take a long detour around the edge of the fence.

Zusammenfassung — Kriterien

Kriterium	Beschreibung
Faktiv-ausdrücke	Ausdrücke, die objektiv-faktische Aussagen anzeigen, sind wichtiger als solche, die subjektiv-kontrafaktische Meinungen andeuten. *Beispiele für objektiv-faktive Ausdrücke*: Wissen anzeigende Ausdrücke ("know that", "it's a fact that", "it is documented that") *Beispiele für subjektiv-kontrafaktische Ausdrücke*: alle Pronomen der 1./2. Person ("I", "you" etc.), den Konjunktiv anzeigende Ausdrücke ("could"/"should"/"would", "seem to", "may[be]"), die persönliche Einstellung anzeigende Ausdrücke ("wish"/"guess"/"think"/"believe", "alas")

Zusammenfassung – Kriterien

* Beispieltext mit subjektiv-kontrafaktischen Ausdrücken (rot):

> The field of reptile cognition is in its infancy, but it already suggests that "intelligence" may be more widely distributed through the animal kingdom than had been imagined.
>
> "They seem to learn the quickest at body temperatures that are very uncomfortable for us," Dr. Burghardt said.

* Beispieltext mit objektiv-faktiven Ausdrücken (grün):

> This flexibility […] has been well documented in birds and primates, and scientists are now beginning to believe that it exists in reptiles, too.
>
> Other studies have documented similar levels of flexibility and problem solving.

Zusammenfassung — Kriterien

- Einige Kriterien sind einander widerstreitend und nur in einer Gesamtbewertung sinnvoll zu verrechnen:
  - Bspw. sind Pronomen zwar kohäsions- bzw. kohärenzstiftend, aber ohne konkret ermittelte Bezugsinhaltswörter nur bedeutungslose Funktionswörter;
  - "well" ist zwar ein Adverb und damit grundsätzlich ein Inhaltswort, jedoch wird es oftmals als bedeutungsloses Füllwort verwendet (vor allem in niedergeschriebenen gesprochensprachlichen Ausdrücken wie z. B. Zitaten).

  Die Gesamtbewertung muss daher einen Ausgleich zwischen verschiedenen Charakteristika eines Ausdrucks herstellen:
  - durch Gewichtung der einzelnen Aspekte;
  - durch Addition vs. Subtraktion positiver vs. negativer Aspekte.

Zusammenfassung — Bewertung

- Die Berechnung einer Gesamtbewertung für die einzelnen Texteinheiten wie Sätze und Absätze (bei extraktivem Summarizing) erfolgt durch 'Scoring' und 'Ranking' der Kriterien:
  - Summe aller Bonus- abzüglich Malusausdrücke (Faktivausdrücke evtl. als Bonusausdrücke);
  - Relationiertheit einer Einheit zu anderen Einheiten (Anzahl und Stärke thematischer Ketten);
  - Topikalität, Frequenz, Informativität von Einheiten usw.

  Die Be- und Verrechnung der einzelnen Zahlenwerte ist je nach Kriterium gesondert festzulegen.

Zusammenfassung — Bewertung

Aufbauend auf dem Scoring-Wert für jede Text-einheit wird eine sortierte Liste pro einzelnem Kriterium erzeugt:

- Die Einheiten werden gemäß ihrer Bewertung für das jeweilige Kriterium absteigend aufgelistet (von 'gut'/ 'viel' nach 'schlecht'/'wenig');
- aus den Einzelrankings pro Kriterium wird ein Gesamt-ranking erstellt, das aus den Rankingpositionen jeder Texteinheit ein gemitteltes Gesamtranking erzeugt:
  - Jedes Kriterium wird dabei gleich gewichtet;
  - einzelne 'wichtige' Kriterien werden höher gewichtet.

Zusammenfassung — Bewertung

Fiktives Beispiel für Gesamtbewertung anhand von sieben Kriterien (S_k = Satz k im Text):

Rang	Krit. 1	Krit. 2	Krit. 3	Krit. 4	Krit. 5	Krit. 6	Krit. 7
0	S_5	S_4	S_3	S_2	S_1	S_2	S_6
1	S_2	S_3	S_4	S_3	S_5	S_8	S_2
2	S_3	S_1	S_0	S_4	S_3	S_3	S_0
3	S_1	S_0	S_5	S_1	S_7	S_4	S_3
…	…	…	…	…	…	…	…

Durchschnittsrang S_3: (2+1+0+1+2+2+3)/7 = 11/7 = 1.57
(für alle anderen Sätze analog)

Zusammenfassung — Evaluation

- Zur Bestimmung der Güte eines extraktiven Summarys sind zwei Schritte notwendig:
  - Erstellen einer Menge von Referenzsummarys aus einer Menge von Volltexten (Korpus), die manuell (intellektuell) von mehreren Personen extraktiv zusammengefasst werden (s. Details unten);
  - Vergleichen der manuellen Referenzextracts mit den entsprechenden maschinell erzeugten Extracts durch verschiedene Evaluationsmaße (s. Details unten).

Zusammenfassung — Evaluation

- Die manuell von Personen erzeugten Referenz-extrakte werden aufgrund implizit-intuitiver Relevanz-Einschätzung von Texteinheiten gewonnen (Textpassagen wie Sätze/Absätze):
  - Auswahl interessant-informativer und/oder repräsentativ-indikativer Einheiten aus dem Text;
  - die intellektuell gewählten Einheiten müssen durch ein Summarizing-System anhand verschiedener berechenbarer Textkriterien möglichst getreu repliziert werden (s. Textkriterien oben).

Zusammenfassung — Evaluation

Manuelles Referenzextrakt des Beispieltextes:

[1] Humans have no exclusive claim on intelligence.
[9] Few scientists bothered to peer into the reptile mind, and those who did were largely unimpressed.
[11] "Certainly in the past, people didn't really think too much of their intelligence. They were thought of as instinct machines."
[12] But now that is beginning to change, thanks to a growing interest in "coldblooded cognition" and recent studies revealing that reptile brains are not as primitive as we imagined.
[36] Dr. Burghardt, for instance, presented monitor lizards with an utterly unfamiliar apparatus, a clear plastic tube with two hinged doors and several live mice inside.
[37] The lizards rapidly figured out how to rotate the tube and open the doors to capture the prey.

Zusammenfassung — Evaluation

- Um ein möglichst 'objektives' oder 'ideales' Extrakt zu erhalten, sollte jeder Text von mehreren Test-personen extraktiv zusammengefasst werden;
- die gewählten Texteinheiten werden nach ihrer Auswahlhäufigkeit zu einem Referenzextrakt vereint:
  - Je häufiger eine Einheit gewählt wurde, desto relevanter ist sie für die Mehrheit der Personen (selbst wenn eine Einheit nur von einer einzigen Person selektiert wurde, ist diese relevanter als eine überhaupt nicht selektierte);
  - nicht ausgewählte Einheiten mit Häufigkeit 0 scheinen objektiv keinerlei informationellen Wert zu besitzen.

Zusammenfassung — Evaluation

- Zur Bestimmung der Güte einer maschinell erzeugten Extrakt-Zusammenfassung werden verschiedene *Evaluationsmaße* verwendet:
  - Das maschinelle Extrakt wird dabei mit dem manuell erzeugten Referenz-/Idealextrakt hinsichtlich des Übereinstimmungsgrades an Texteinheiten verglichen (z. B. Überlappungsgrad von Sätzen);
  - je mehr Einheiten das maschinelle Extrakt mit dem manuellen Referenzextrakt gemeinsam hat, desto besser ist die automatische Zusammenfassung.

Zusammenfassung — Evaluation

Gängige Evaluationsmaße für Extract-Summarys:

- Einzelmaße:
  - *Recall*: Anzahl *übereinstimmender* Einheiten / Anzahl *relevanter* Einheiten im Referenzextrakt;
  - *Precision*: Anzahl *übereinstimmender* Einheiten / Anzahl *ermittelter* Einheiten im Systemextrakt.

Übereinstimmende Einheiten treten sowohl im manuellen Referenz- als auch maschinellen Systemextrakt auf.

Relevante Einheiten sind alle *manuell* gewählten Einheiten des *Referenzextrakts*, ermittelte Einheiten sind alle vom *System* ausgegebenen, für *relevant befundenen* Einheiten.

Zusammenfassung — Evaluation

- Kombinationsmaße:
  - *R-Preci-sion*: Anzahl *übereinstimmender* Einheiten / Anzahl *ermittelter* = *relevanter* Einheiten.
    - Sollen vom System genau so viele Einheiten ermittelt werden, wie im Referenzextrakt vorhanden sind, dann unterscheiden sich Recall und Precision nicht mehr (man hätte das Maß entsprechend auch R-Recall nennen können);
    - das Maß gibt normierte Werte zwischen 0 und 1 aus, wobei höhere Werte gegen 1 ein besseres maschinelles Extrakt anzeigen.
  - *F-Mea-sure*: (2·Recall·Precision) / (Recall+Precision).

    F-Measure liefert eine zwischen Recall und Precision gleichgewichtete Gesamtaussage, die *ein* Leistungsmaß für einen Summarizer ausgibt.

Zusammenfassung — Evaluation

Recall und Precision machen generell zwei verschiedene Aussagen über die Leistung eines Summarizing-Systems:

- *Fokus auf Umfang*: Durch Recall wird ausgedrückt, wie gut das System in der Lage war, *alle tatsächlich relevanten* Einheiten zu ermitteln, d. h. nur solche Einheiten, die die Nutzer auch für relevant halten (d. h. alle Personen, die stellvertretend an den Referenz-Extrakten beteiligt waren);
- *Fokus auf Genauigkeit*: Durch Precision wird ausgedrückt, wie viele der vom System ermittelten/ausgegebenen Einheiten *für die Nutzer tatsächlich relevant* sind (wobei nicht-ermittelte Einheiten nicht bewertet werden).

Zusammenfassung — Evaluation

Beispielberechnungen aller Maße anhand obigen Referenzextrakts:

- Annahmen:
  - Im idealen Referenzextrakt befinden sich die sechs relevanten Sätze [1], [9], [11], [12], [36], [37] (s. oben);
  - im Systemextrakt befinden sich die sieben ermittelten Sätze [1], [2], [10], [11], [36], [38], [39] (wobei [1] nur der siebtbeste sein soll).
- Vorberechnungen:
  - Anzahl übereinstimmender Sätze: 3 (nämlich [1], [11], [36]);
  - Anzahl relevanter Einheiten im Referenzextrakt: 6;
  - Anzahl ermittelter Einheiten im Systemextrakt: 7
    (das System 'kennt' die Anzahl tatsächlich relevanter Einheiten nicht, solange diese Vorgabe nicht explizit vom Nutzer gemacht wird).

Zusammenfassung — Evaluation

- Konkrete Werte der Evaluationsmaße:
  - Recall = 3 / 6 = 0.50:
    Die Hälfte aller als relevant befundenen Sätze wurde vom System auch ermittelt.
  - Precision = 3 / 7 = 0.43:
    Drei von Sieben ermittelten/ausgegebenen Einheiten sind auch relevant (d. h. die anderen sind irrelevant und damit gewissermaßen 'Informationsschmutz').
  - R-Precision = 2 / 6 = 0.33:
    Da nur 6 statt 7 Sätze vom System ermittelt werden sollen und Satz [1] nur der siebtbeste für das System ist, fällt dieser aus der Betrachtung der übereinstimmenden Sätze heraus.
  - F-Measure = (2·0.50·0.43) / (0.5+0.43) = 0.46.

Zusammenfassung — Herausforderungen

- Herausforderungen des automatischen Zusammenfassens:
  - Auflösung pronominaler Bezüge:
    - Jedes Pronomen steht für einen bereits im Text zuvor erwähnten Ausdruck (z. B. "he" für "Einstein", "it" für "theory of relativity");
    - Pronomen rechnen zu den Funktionswörtern und werden daher im Textanalyseprozess eliminiert, stehen jedoch eigentlich stellvertretend für wichtige und inhaltsvolle Konzepte im Text;
    - daher sollten Pronomen besser durch die Ausdrücke ersetzt werden, für die sie stehen;
    - dies ist jedoch nur sehr aufwändig zu realisieren.

Zusammenfassung — Herausforderungen

- Wiedergabe des Textinhalts mit eigenen Formulierungen:
  - Das Abstracting von Texten ist im Gegensatz zum Extracting eine wesentlich anspruchsvollere Aufgabe, da der Text in eigenen Worten wiedergegeben werden soll;
  - der Text muss daher vorher semantisch analysiert und intern repräsentiert werden, d. h. der Text muss bis zu einem gewissen Grad auch 'verstanden' werden;
  - aus der internen Repräsentation ist durch eine Texterzeugungskomponente ein lexikalisch und grammatikalisch prägnant und präzise reformulierter Summary-Text zu erzeugen.

Übung – Vertiefung

- Ermitteln Sie aus folgendem Text die Ihrer Meinung nach wichtigsten Sätze und stellen Sie diese zu einem extraktiven Summary zusammen. Begründen Sie Ihre Auswahl!

[0] Your Initials May Influence Your Job

[1] The initials of your name may influence where you choose to work, new research suggests.

[2] While it sounds like a joke, a well-known psychological theory called the name-letter effect maintains that a person's behavior may be influenced by his or her name.

Übung – Vertiefung

[3] As my colleague Stephanie Rosenbloom reported earlier this year, "people like the letters in their own names (particularly their initials) better than other letters of the alphabet."

[4] Johnsons are more likely to wed Johnsons, women named Virginia are more likely to live in (and move to) Virginia, and people whose surname is Lane tend to have addresses that include the word "lane," not "street."
[5] During the 2000 presidential campaign, people whose surnames began with B were more likely to contribute to George Bush, while those whose surnames began with G were more likely to contribute to Al Gore.

[6] Researchers from Ghent University in Belgium decided to test the "name-letter effect" to determine if it is powerful enough to influence a person's place of employment.

Übung – Vertiefung

[7] The psychologists analyzed a database containing information about Belgian employees who work full-time, looking at the employees' names and how often the first initial matched the first letter of their company's name.

[8] While a certain number of matches would be expected by chance, the researchers found that there were 12 percent more matches than was expected based on probability estimates.
[9] The findings, published in Psychological Science, showed that for about one in nine people whose initials matched their company's initial, choice of employer seems to have been influenced by the fact that the letters matched.

[10] The authors concluded that they "have demonstrated that people are more likely to work for companies with initials matching their own than to work for companies with other initials."

Übung – Vertiefung

[11] I have personally always been skeptical of the theory, but also confess that as someone with the initials TLP, I have a surprising number of examples of T's, L's and P's in my life.
[12] (And now I do work at The Times!)

[13] What do you think?
[14] Is it silly psychology, or have you seen any evidence of the name-letter effect in your life?

well.blogs.nytimes.com/2008/10/23/how-your-name-mayinfluence-your-behavior/

Selektierte Sätze:

#	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]
+/–															

Information-Retrieval

- Information-Retrieval bezeichnet den Vorgang, auf eine (Suchan-)Frage eine Antwort in Form von thematisch relevanten Dokumenten zu finden:
  - Die Frage wird als Menge natürlichsprachiger Suchterme angegeben, die das *subjektive Informationsbedürfnis* des Nutzers widerspiegeln (im Gegensatz zum *objektiven Informationsbedarf*, der nicht feststellbar ist);
  - die Antwort erfolgt als Menge von Dokumenten oder Textpassagen aus diesen Dokumenten, in denen die gewünschte Information idealerweise enthalten ist.

  Vorausgesetzt ist eine 'Kollektion' von Dokumenten, in denen die potenziellen Antworten enthalten sind.

Information-Retrieval

- Die Frage wird wie die Antwort auch als Textdokument aufgefasst, die aus einer Reihe von Termen bestehen:
  - Sowohl das Fragedokument als auch das potenzielle Antwort-dokument werden zunächst indexiert (z. B. gewichtet nach Häufigkeiten durch das tf·idf-Maß);
  - die Indexterme der beiden Dokumente dienen dann als Basis für einen Ähnlichkeitsvergleich:
    - Je ähnlicher sich Frage- und Antwortdokument(e) hinsichtlich ihres Termbestandes sind, desto besser 'passt' die Antwort auf die Frage;
    - im einfachsten Fall eines Ähnlichkeitsabgleichs wird die Anzahl und Abfolge der überlappenden Terme aus der Frage und der Antwort als Kriterium verwendet.
  - Abschließend erfolgt ein 'Ranking' der Antworten gemäß der Ähnlichkeiten zwischen Frage und Antwortdokumenten.

Literatur

- Hahn, U. (2013): Automatisches Abstracting. In Kuhlen & al. (2013), S. 286–301.
- Kuhlen, R. & Semar, W. & Strauch, D. (^{6}2013; Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. De Gruyter: Berlin.
- Lepsky, K. (2013): Automatische Indexierung. In Kuhlen & al. (2013), S. 272–285.
- Moens, M.-F. (2000): *Automatic Indexing and Abstracting of Document Texts*. Boston u. a.: Kluwer.
- Nohr, H. (^{3}2005): *Grundlagen der automatischen Indexierung*. Berlin: Logos.
- Reischer, J. (2010): *Retrieval und Ranking informativer Textpassagen*. Universität Regensburg: Habilitationsschrift.