Die Anwendung von Morphologieanalyse in Information Retrieval-Systemen

Morphologieanalyse und Lexikonaufbau (10. Vorlesung)

Übersicht

- Was ist Informationslinguistik?
- Masszahlen des Information Retrieval
- Informationslinguistik
- Graphematisch-phonologische Verfahren
- Schreibfehlerkorrektur
- Morphologische Verfahren
- Syntaktische Verfahren
- Semantische Verfahren
- Statistische Verfahren
- Linguistische Probleme in Information Retrieval Systemen

Teilw. nach [Hahn und Sonnenberger 91]: Einführung in die Informationslinguistik. Uni Konstanz.

Was ist Informationslinguistik?

Informationslinguistik untersucht sprachliche Probleme der Textanalyse, wie sie typischerweise im Kontext von Information Retrieval (IR)-Systemen auftreten.

Was ist Information Retrieval (IR)?

Informationsgewinnung aus textuellen Datenbanken, i.a. schlüsselwortbasiert (Deskriptoren). Man klassifiziert IR-Systeme nach:

- Art der zur Verfügung stehenden Information: (Volltexte, Zusammenfassungen)
- Format der Texte (strukturiert: z.B. Bibliographieeinträge; unstrukturiert: z.B. Volltexte)
- nach der Art der Bestimmung der Deskriptoren (von Hand, automatisch)
- nach der Art der Speicherung der Deskriptoren (flache, hierarchische, netzwerkartige Systeme)
- nach der Art des Retrievals (linguistisch basiert, statistisch basiert)

Was ist Indexierung?

Die Abbildung des Inhaltes eines Dokumentes auf eine Menge von relevanten Begriffen.

Genauer: Die Zuordnung von Deskriptoren und Notationen zu Dokumenten zwecks ihrer inhaltlichen Erschliessung und gezielten Wiederauffindung (vgl. DIN 31 623).

Vgl. Heinz-Dirk Luckhardt (Universität des Saarlandes): Automatische und intellektuelle Indexierung

IR und seine nahen Verwandten im Vergleich

IR Probleme

◊ dasselbe Wort kann verschiedene Typen von Objekten bezeichnen, wie auch umgekehrt.

◊ Wörter liegen selten in ihrer Grundform im Text vor, sondern häufiger in einer flektierten Form. Die meisten IR-Verfahren verwenden zu wenig morphologische Informationen → Verlust an *Recall (Ausbeute)* (viele der relevanten Dokumente werden nicht gefunden).

◊ Schlüsselwortsuche missachtet jede Form der syntaktischen und letztlich semantischen Zusammenhänge innerhalb eines Satzes → Verlust an *Präzision* (viele der gefundenen Dokumente sind nicht relevant).

◊ Nur ganze Dokumente werden gefunden, oft sehr viele → zeitaufwendiges manuelles Durchlesen

Alternativen zu IR

Informationsextraktion (IE), Textbasiertes Fragenbeantworten, Antwortextraktion (AE, Q&A)

Informationsextraktion (IE)

Informationsextraktion (IE) erlaubt das effiziente Absuchen grosser Textmengen auf präzise, vordefinierte Fragestellungen hin, die einen Sachverhalt ausdrücken (Message Understanding Conference, http://www.muc.saic.com). Da dabei ein fixes, datenbankähnliches Informationsraster gefüllt wird, können nur sehr eng definierte thematische Bereiche abgedeckt werden.

Textbasiertes Fragenbeantworten

Der ideale Lösungsansatz wäre zweifellos der Einsatz von Systemen zur automatischen Fragenbeantwortung über Texten. Die Erfahrungen bei der Entwicklung derartiger Systeme, z.B. LILOG (Herzog 1991) haben allerdings gezeigt, dass der Entwicklungsaufwand für derartige Systeme sehr gross ist.

Answer Extraction (AE, Q&A)

Geht davon aus, dass man in den vorhandenen Texten oft Stellen lokalisieren kann, welche die *Antwort auf eine Frage explizit enthalten*. Im Unterschied zum IR baut AE auf einer vollständigen morphologischen und syntaktischen Analyse auf. Einige Aspekte der Semantik wie thematische oder terminologische Relationen werden analysiert. Anderes Szenario: Benutzeranfragen, Problemlösen. Q&A-Track der TREC-8 (1999).

Masszahlen des Information Retrieval

Die Qualität eines Information Retrieval Vorgangs wird durch zwei Masszahlen (Recall und Precision) beschrieben, die auf folgenden Parametern beruhen:

- Anzahl der gefundenen relevanten Dokumente: `F`
- Anzahl aller relevanten Dokumente: `R`
- Anzahl aller gefundenen Dokumente: `A`

Recall (Vollständigkeit der Suche, "Ausbeute")

```
Recall    =  F/R
```

Precision (Genauigkeit der Suche)

```
Precision =  F/A
```

Merke: 'Relevanz' ist das Mass der Übereinstimmung zwischen einem Dokument und der Suchanfrage aus der Sicht eines Experten.

F-Wert (Kombination von Precision und Recall)

$$\text{F-Wert} = \frac{2}{\frac{\text{Recall} + \text{Precision}}{\text{Recall} * \text{Precision}}}$$

'Relevanz' ist das Mass der Übereinstimmung zwischen einem Dokument und der Suchanfrage aus der Sicht eines Experten.

Informationslinguistik

Die sprachlichen Probleme der Informationslinguistik betreffen sämtliche Ebenen der sprachwissenschaftlichen Betrachtung.

1. Graphematisch-phonologische Verfahren

1.1 Erkennung von lautlichen oder Schreibvarianten

- Personennamen in Presse-Datenbanken, Personal-, Patienten- oder Kunden-Dateien

```
MEIER <=> MEYER <=> MAIER <=> MAYER <=> MAYR
GORBACHOW <=> GORBACHEV <=> ...
GHADDAFY <=> KHADAFY
```

- geographischer Name

```
MUENCHEN <=> MUNICH <=> MONACO DI BAVIERA <=> ...
```

- Produktnamen in Patentämtern

```
CRONAT <=> SOMAT
```

- Abkürzungen / Akronyme

```
CO2  <=> Kohlenstoffdioxyd
NATO <=> North Atlantic Treaty Organisation
```

1.2 Schreibfehler-Erkennung und -Korrektur

Untersuchungen (in den 80er Jahren) haben ergeben, dass in On-line Datenbanken teilweise mehr als 10% Schreibfehler vorkommen (d.h. jedes 10. Wort ist falsch geschrieben).

80% der Schreibfehler lassen sich auf die folgenden 4 Fehlertypen zurückführen:

```
Auslassung              CHMICAL
Einfügung              CHEMEICAL
Substitution           CHEMECAL
Vertauschung           CHMEICAL
               ==> CHEMICAL
```

Anzahl der möglichen Schreibfehler (Einfachfehler) in einem Wort der Länge `n` (Ausgangsbasis 26 Buchstaben, Bindestrich, Hochkomma)

```
Auslassung          n
Einfügung           28 * (n + 1)
Substitution        28 * n
Vertauschung        n - 1
```

oder: Wieviele Buchstaben stimmen überein? (Reihenfolge?)

Zusammen mit Trunkierung kann 'Fehlerkorrektur' auch als Lemmatisierungshilfe fürs IR missbraucht werden.

Wortlistenabhängige Verfahren zur Schreibfehlererkennung:

Abgleich mit einer Wortliste (mit oder ohne Lemmatisierung)

Problem: Wenn der Schreibfehler ein anderes korrektes Wort ergibt, wird er nicht erkannt.

Schreibfehlererkennung kann in IR (zus. mit Trunker) primitive Lemmatisierung ermöglichen.

Wortlistenunabhängiges Verfahren zur Schreibfehlererkennung:

N-Gramm-Analyse: basiert auf der Untersuchung der Häufigkeit von Buchstabenfolgen einer bestimmten Länge (meist Länge n=2 oder n=3).

Anzahl möglicher n-Gramme: (angenommen 28 Zeichen im Alphabet)

Bigramme: Trigramme:

$28^2 = 784$ $28^3 = 21'952$

In einem grösseren Textkorpus treten ca. 70% der möglichen Digramme und ca. 25% der möglichen Trigramme auf.

Bsp.: *Cmputer* wird als Fehler erkannt, da Trigramm *cmp* im Deutschen nicht vorkommt.

2. Morphologische Verfahren

- Erkennung und Reduktion von Flexionsvarianten einer Grundform

```
COMPUTER
COMPUTER'S
COMPUTERS
COMPUTERS'
==> COMPUTER
```

- Erkennung und Reduktion von Derivationsvarianten einer Stammform

```
COMPUTER
COMPUTATIONAL
COMPUTED
==> COMPUT(E)
```

- Erkennung von Bestandteilen eines Kompositums (bes. bei Fugenlaut und Ellision eines Konsonanten)

```
Zeitungsartikel
==>  Zeitung + Artikel
Schrittempo
==>  Schritt + Tempo
```

Idealerweise mittels professionellen Morphologieanalyseprogrammes, siehe z.B. Two-level-Morphology.

Mit geringerem Aufwand sind folgende Annäherungen an Lemmatisierung verbunden:

- Trunkierung: Die letzten (z.B. 3) Buchstaben werden abgetrennt
  Bsp.:*multinationa(ler), Speicher(ung), f(rei), Compu(ter), ge(hen)*
- Endungslexikon: Mögliche Endungen werden abgetrennt
  Bsp.: *Carbide coatings on graphite fibers by liquid metal transfer agent method*
- 'Fehlerkorrektur', z.B. zusammen mit (Endungs)lexikon
  Bsp.: *W<u>ä</u>ld-er, Z<u>ö</u>g-ling, kopie<u>r</u>-te, Währung<u>s</u>-union*

- Erkennung von Eigennamen und Termen (gesondert zu behandelnde Indexterme)

Lexikalisch: Eigennamenlexikon oder Negativverdacht: N nicht in allg. Lexikon → Eigenname-Kandidat

Kollokationsforschung: Begleiter (Frau, Dr., AG, Ltd.), Mehrwortterme (statistisch)

3. Syntaktische Verfahren

3.1 Erkennung von komplexen (mehrgliedrigen) Nominalphrasen

Beispiel:

```
EIGENVALUE PROBLEM
INFORMATION THEORY
DEDUCTIVE DATA BASE
```

Im Information Retrieval werden dafür besonders Abstandsoperatoren (`Adjacency') verwendet.

Als informationslinguistische Lösungsansätze kommen folgende Verfahren in Betracht:

- vollständige Sytaxanalyse (Parsing)
- partielle Syntaxanalyse
  - wörterbuchunabhängig (Begrenzerverfahren, Chunking)
  - wörterbuchabhängig (Bottom-Up Chart-Parsing, Island Parsing)

Wörterbuchunabhängige Syntaxanalyse basiert auf der Segmentierung eines Textes über die Funktionswörter (Artikel, Präpositionen, Konjunktionen, Determiner-Pronomen) und Interpunktion. Diese werden interpretiert als Begrenzer, die eine Nominalgruppe einleiten oder abschliessen. Eine Verfeinerung des Verfahrens ist möglich über die Ermittlung der statistischen Relevanz von Begrenzerpaaren.

Fast noch geläufiger ist der Einsatz eines Taggers, auf dessem Output sich tag-basierte reguläre Ausdrücke anwenden lassen, z.B. (/ART)? (/ADJA)* /NN für einfache, ungeschachtelte NPs.

Beispiele:

```
The phosphoric acidity has …
was generally controlled by …
the porosity formed by …

Bei [der/ART hier/ADV vorgelegten/ADJA chemischen/ADJA Analysemethode/NN] hat sich
gezeigt, …
```

3.2 Erkennung von nominalen syntaktischen Paraphrasen

```
WATER TREATMENT  <=> TREATMENT WITH WATER
NEUTRON EXCHANGE <=> EXCHANGE OF NEUTRONS
```

3.3 Erkennung und Auflösung von Nominalkomposita

```
EIGENWERTBERECHNUNG  ==> BERECHNUNG, EIGENWERT
PROGRAMMENTWURF  ==> PROGRAMM, ENTWURF
```

3.4 Erkennung von attributiv expandierten Varianten von Nominalphrasen

```
BERECHNUNG VON EIGENWERTEN
BERECHNUNG EINFACHER EIGENWERTE
BERECHNUNG DICHT BENACHBARTER EIGENWERTE
BERECHNUNG ZWEIER ISOLIERTER EINFACHER EIGENWERTE
==> BERECHNUNG, EIGENWERT
```

4. Semantische Verfahren

- Extraktion inhaltlich signifikanter Terme durch Abgleich mit Stoppwort- oder Positiv-Listen (Stoppwort = nicht-referierendes Wort aus geschlossener Klasse, eingetragen in Liste)
- Bestimmung inhaltlich signifikanter Terme für Indexing und Retrieval durch Berechnung von einfachen Termgewichten
- Thesaurus-basierte Verfahren (hand- oder maschinenerstelltes Netz oder Hierarchie)

  Beispiel: Wordnet

5. Statistische Verfahren

- Tagging: Inhaltsworttags und Funktionsworttags statt Stoppwortlisten.
- Term Frequency * Inverse Document Frequency (tfidf-Verfahren): Nur Inhaltswörter über einem gewissen tfidf-Wert sind Indexterme eines gegebenen Dokumentes.
- Chi-square-Verfahren: Nur Inhaltswörter unter einem gewissen Chi-square-Wert (im Vergleich zur Gesamtmenge an Dokumenten) sind Indexterme eines gegebenen Dokumentes.
- Kollokationen als Mehrwortterme, Eigennamenerkennungsmethoden
- Word Sense Disambiguation (WSD) und -Clustering aufgrund des Kontextes.
  - ◊ Ganz andere Kontexte → polysemisches Wort: (Bank, Garten) vs. (Bank, Geld)
  - ◊ Objekte, die gleich häufig von den selben Verben regiert werden, sind wohl semantisch nah verwandt.

Linguistische Probleme in Information Retrieval Systemen

(nach [Kuhlen 86]: Informationslinguistik.)

- Eliminierung von bedeutungslosen Wörtern (Stoppwortlisten)
- Reduktion von Wortformen auf Grund- oder Stammformen (Lemmatisierung)
- Wortzerlegung in kleinere semantische Einheiten (z.B. in Morpheme)
- Bestimmung der Wortart (Tagging)
- Partielles Parsing, um relevante Nominalphrasen und Präpositionalphrasen zu erkennen
- Paraphrasen für komplexe Nominalphrasen
- Partielles Parsing, um einfache syntaktische Relationen zu erkennen (Subjekt, Objekt, Präpositionalobjekt)
- Erkennung von Topik (Thema) und Comment (Rhema)

- Disambiguierung von Homographen mit Hilfe von Kontext oder Parsing

```
Wald+Bäume+Kiefer  => Baum
Zahnarzt+Schmerzen+Kiefer  => Kopfteil
```

- Elementares Parsing zur Erkennung von syntaktischen Einheiten (Teilsätze; z.B. Relativsätze oder elliptische Koordination abgrenzen und auflösen)
- Zusammenstellung von Wortgruppen über assoziative Relationen (-> z.B. über einen Thesaurus)
- Zusammenstellung von Verbrahmen (aufgrund von Valenzangaben)
- Erzeugung von Prädikat-Argument Strukturen (Prädikatenlogik)
- Integration von gewichteten semantischen Relationen

Gerold Schneider
Date of last modification: June 12, 2001
Source: http://www.ifi.unizh.ch/CL/gschneid/LexMorphVorl/Lexikon10.IR.html

Ruge & Goeser 98:

Erwartungsgemäß führt Linguistik zu verbesserter IR-Leistung, in der Praxis ist das aber nur teilweise zu beobachten.

- Stemming: Krovetz (1993) +, Harman (1991) -. Tendenz + (Paice 1994). Überdifferenzierung, Überanalyse, sg/pl
- Synonyme: Rada (1989) +, Kristensen (1990) +, Voorhees (1994) -. Domänenspezifischer Thesaurus wichtig. Auch rel. schlechte Syntaxanalysen sinnvoll.
- Polysemie, WSD: Ein IN-Thema! Voorhees (1993) -, Sanderson (1994) -. Abhängig von der Güte der WSD. 10% oder 30% Schwelle? Bei guter WSD deutliche Verbesserung.
- Phrasen, Syntax: Smeaton (1995) -, Strzalkowski (1995) +. Phrasen sind seltener als Einzelwörter, Multiwortterme (*schneller Brüter*) vs. synt. Modifikation (*schneller Flug*)
- Bei kurzen Anfragen hilft Linguistik mehr als bei langen. Benachteiligung durch TREC Standardevaluierung (http://trec.nist.gov/)