Computerlinguistik in Information und Dokumentation

Kurs für wissenschaftliche Bibliothekare

Teil I: Linguistische Ebenen, computerlinguistische Anwendungen und Ansätze

Simon Clematide
Institut für Computerlinguistik
Universität Zürich

9. März 2006

http://www.cl.unizh.ch/siclemat/talks/zb/

Vorspann: Was ist Computerlinguistik?

Wissenschaftliche Disziplin

- Entwicklung von Theorien und Ansätzen, welche der maschinellen Verarbeitung von **natürlicher Sprache** dienen
- Verknüpfung von Linguistik, Informatik, Kognitionswissenschaften

Praktische Disziplin

- Entwicklung von lauffähigen Systemen, welche Information in Form natürlicher Sprache verarbeiten können
  - Automatische Übersetzung, intelligente Informationsextraktion und -wiedergewinnung, natürlichsprachliche Benutzerschnittstellen

Das Programm

Übersicht

- **Linguistische Ebenen** von geschriebener Sprache und **computerlinguistische Anwendungen** und **Ansätze**
  - Morphologische Ebene (F4)
    - Analyse, Reduktion auf Grundformen
  - Ebene der Wörter (F8)
    - Wortartenbestimmung, Lexikalische Semantik, Termini, Eigennamen
  - Ebene der Syntax (F13)
    - Beziehung von Syntax und Semantik
  - Ebene des Texts (F16)
    - Textverstehen und Fragebeantwortung
- Maschinelle Übersetzung (F18)
- Textzusammenfassung (F23)

Morphologische Ebene

Womit befasst sich die Morphologie?

Wortstruktur und ***Wortbildung***!

- Flexion

  such+en, such+e, such+test, such+ten, ge+such+t, such+end,…

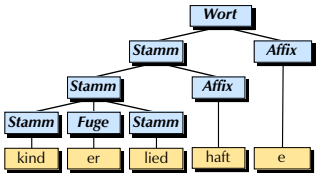
  Frucht, Früchte
- Derivation

  suchen, Suche

  Frucht, frucht+en, frucht+bar, un+frucht+bar, Un+frucht+bar+keit,…
- Komposition

  Such+ergebnis,

  Text+zusammenfassung+s+system



Beispiele morphologischer Analyse

Hierarchische Wortbildungsanalyse von CANOO

Flache Wortstrukturanalyse von GERTWOL

```
Suchergebnisses
"<*suchergebnisses>"
        "*such#er|geb~nis"   S NEUTR SG GEN
```


Probleme morphologischer Analyse

Kategorielle Mehrdeutigkeit

- "Müller" als Eigenname oder Substantiv

Strukturelle Mehrdeutigkeit

- Unterschiedliche Analysen bzw. unklare Gruppierung der Bestandteile

Unvollständigkeit

- Neubildungen, Spontanbildungen, Fremdwörter, …
- **Zielkonflikt**
  Je vollständiger, umso mehrdeutiger!

Überanalyse

- Kein Unterschied zwischen lexikalisierter Form und produktiver Bildung!

```
"<*müller>"
        "*müller"  S EIGEN Famname SG NOM
        "*müller"  S MASK SG NOM
...
"<*verbrechen>"
        "*verb#rechen"  S MASK SG NOM
        "*verb#rech~en"  S NEUTR SG NOM
        "*ver|brech~en"  S NEUTR PL DAT
        "ver|brech~en"  * V INF
        "ver|brech~en"  * V IND PRƒS PL1
...
"<eine>"
        "ein"  ART INDEF SG NOM FEM
        "ein"  ART INDEF SG AKK FEM
        "einer"  PRON INDEF SG NOM FEM
        "einer"  PRON INDEF SG AKK FEM
        "ein~en"  V IND PRÄS SG1
        "ein~en"  V KONJ PRÄS SG1
        "ein~en"  V KONJ PRÄS SG3
        "ein~en"  V IMP PRÄS SG2
...
"<*erdbeere>"
        "*erd#beere"  S FEM SG NOM
...
```

Analyseauszüge von GERTWOL

Nützlichkeit morphologischer Analyse

Grundformenbestimmung flektierter Formen

- Linguistisch fundierte Normalisierung
  - Im Gegensatz zum blossen Abschneiden (Trunkierung) oder heuristischen Reduzieren (Stemming)

Mütter → Mutter

Wortableitung auflösen

- Assoziieren von strukturell verwandten Begriffen
  - Interessant sind Ableitungen, welche nur Wortartwechsel beinhalten!

baulich → Bau
Linguistik → linguistisch

Dekomponierung/Ergänzung von Komposita

- Insbesondere Neu- bzw. Spontanbildungen, deren Bedeutung sich aus den Einzelteilen ergibt.
  - Interessant sind Teile, welche im Gebiet belegt sind!

Wohnbauförderungsmöglichkeiten → Wohnbauförderung

Text- und Diskurstheorie → Texttheorie

Ebene der Wörter

Bestimmung von Wortarten im Text

- Kategoriell mehrdeutige Wortformen können in ihrer textuellen Umgebung zuverlässig (>90%) desambiguiert werden.

Faktoren

- Vorkommenshäufigkeit der einzelnen Wort-Kategorie-Paare
- Kombinationsmöglichkeiten im Satz

Einsatz

- Erfolgreiche und nützliche Grundtechnik in vielen Bereichen der Sprachverarbeitung
  - Grundformerkennung, Terminologieerkennung, syntaktischer Analyse, etc.

Lexikalische Semantik

Wie lässt sich die Bedeutung eines Worts angeben?

- Klassisch Charakterisierung: Umschreibung, Definition

Relationale lexikalische Semantik = Bedeutungsbeziehungen

- Durch Angabe von Synonymen, Hypernymen, Hyponymen, Antonymen usw., welche ein **Netz** von verknüpften Bedeutungen ergeben
- Wortnetze für viele Einzelsprachen (engl. "WordNet", dt. "GermaNet")
  - $Bank_1$ ist synonym zu Geldinstitut, Kasse, Geldhaus und hyponym zu "wirtschaftliche Institution"
  - $Bank_2$ ist hyponym zu "Sitzmöbel"

Bedeutungsdesambiguierung im Kontext

- "Bank" bedeutet $Bank_1$, wenn in der Text-Umgebung Wörter mit Bedeutungen aus dem Gebiet Geldwesen vorkommen.
- "Bank" bedeutet $Bank_2$, wenn …

Terminologische Ebene

Spezialsprachliche Fachbegriffe

- Oft komplexe Struktur
  - Zusammensetzung
  - Mehrteilige Wörter

Computer
elektronischer Rechenanlage
free indexing
unendliche Reihe

Aufgaben

- Identifizieren von (ev. leicht variierten) Fachbegriffen
  - Benutzung von Begriffsvarianten ist extrem verbreitet
- Zuordnung in Fachgebiet
  - Oft über Klassifikation von Dokumenten
- Erkennung/Erschliessung von neuer Terminologie in Texten
  - Automatisches Extrahieren von Terminologiekandidaten
  - Terminologie findet sich in Nominalphrasen

Ansätze der Terminologieerkennung

Linguistische Methoden

- Sprachspezifische Wortgruppenmuster für Nominalphrasen

Adjektiv + Nomen
Nomen + Nomen
Nomen + »of«-Präposition + Adjektiv + Nomen

non-financial enterprise
interbank market
settlement of cross-border payments

Statistische Methoden

- Fachbegriffe kommen in Fachtexten übermässig häufig vor.
- Teile von mehrteiligen Fachbegriffe kommen übermässig häufig nur gemeinsam vor.

Aber

- Linguistische und statistische Kriterien für Termhaftigkeit sind immer heuristisch.
- Welche Termini sind relevant? Abhänging von Verwendungszweck (normativ, deskriptiv) und Zielpublikum.

Eigennamen – "Named Entities"

Uninteressant für Linguistik –

Vital für praktische Systeme

- Erkennung von Personennamen und Organisationen

Lise Meitner an **Otto Hahn**: Briefe aus den Jahren 1912 bis 1924

- Meistens Verwendung von Listen und Mustern mit Kontexteinschränkungen

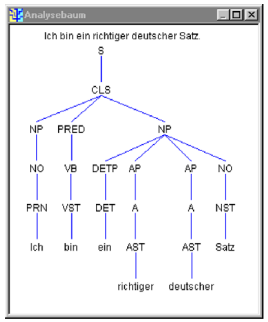
VORNAME + <Grossgeschriebenes Wort>

- Einfache semantische Desambiguierung
  - "Hahn" hat hier nichts mit Geflügel oder Sanitärinstallation zu tun!
- Oft auch als Erkennung von Datumsangaben, Zahlausdrücke, Währungsangaben, ...

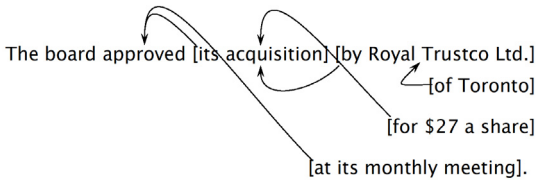
Die syntaktische Ebene

Womit befasst sich die Syntax?

- **Aufbau** von Phrasen und **Abhängigkeit** zwischen Phrasen



Syntaxbaum aus Übersetzungssystem



Komplexe Abhängigkeiten in realen Sätzen

Syntaktische Mehrdeutigkeit

Viele Sätze sind semantisch und/oder syntaktisch mehrdeutig.

Der Mann sah die Frau im Park mit dem Fernrohr.

- Kaum ein Problem für Menschen, da er über viel situationales Hintergrund-, Sprach- und Weltwissen verfügt.
- Riesiges Problem für Computer, welche nur symbolische Grössen auf Grund von Kombinationsregeln verrechnen!

Variabilität und Ähnlichkeit

Der gleiche Sachverhalt in unterschiedlichem Kleid

Utilisation de vues aériennes et inventaire complet des dégât
Inventaire des dégât causés par les tempêtes au moyen de vue aériennes
Inventaire des dégâts causés par les tempêtes à l'aide de vue aériennes

Ähnliche Beschreibung mit unterschiedlichem Sinn

design computer vs. computer design
Export von Autos aus Deutschland nach den USA
Export von Autos aus den USA nach Deutschland

Sprache und Bedeutung

- Bedeutung ergibt sich nicht aus der Summe der verwendeten Wörter.
- Syntaktische Versprachlichung ist nicht durch Bedeutung bestimmt.

Textuelle Ebene

Womit befasst sich die Textlinguistik?

- Satzübergreifende Strukturen und Abhängigkeiten
  - Thema und Kohärenz von Texten und Diskursen
    - Bestimmen von anaphorischen Bezügen
  - Kommunikationsfunktion
  - Weltwissen

F: Wann ist das Hetjensmuseum geöffnet?
A: Von 10 Uhr bis 17 Uhr.

F: Ist es um 14 Uhr geöffnet?
A: Ja.

Informationsgewinnung

Im Palais Nesselrode ist das Hetjensmuseum, das 1909 eröffnet wurde, untergebracht.
Es befindet sich an der Ecke Schulstrasse und Hafengasse.
Die Keramiksammlung umfasst zehntausend Objekte.
Der Eintritt der Ausstellung, die von 10 bis 17 Uhr geöffnet ist, beträgt 2 DM.

Informationstext aus LILOG

Textverstehen und Fragebeantwortung

Projekt LILOG (1986-91)

- **Li**nguistic and **Lo**gic Methods for Machine Understanding of **G**erman
- Ziele
  - Fundierte linguistische Analyse und Wissensrepräsentation
  - Informationsgewinnung durch Anwendung von Logik
- Sprachwissen
  - Anaphorische Bezugsketten
    - "Hetjensmuseum"–"Es";
    - "Keramiksammlung"–"Ausstellung"–"die"
- Weltwissen
  - Falls Ausstellung geöffnet, dann Museum geöffnet
  - Fall von 10 Uhr bis 17 Uhr geöffnet, dann um 14 Uhr geöffnet
- Fazit
  - 60 Personenjahre Entwicklung für einige Seiten Text…

F: Wann ist das Hetjensmuseum geöffnet?
A: Von 10 Uhr bis 17 Uhr.

F: Ist es um 14 Uhr geöffnet?
A: Ja.

Maschinelle Übersetzung

Unterschiedliche Zielsetzungen

- Menschliche Übersetzungsqualität durch vollautomatisches System
- Rohübersetzungen mit kompetenter menschlicher Nachredaktion (ev. Vorredaktion)
  - System integriert Formatierungen, Illustrationen innerhalb der Textverarbeitung
  - System markiert Problemstellen wie unbekannte Wörter
  - Übersetzer stellt System sorgfältig auf Sachgebiet ein (Lexikonpflege)
- Ad-hoc-Übersetzungen zum kurzfristigen Zugänglichmachen des ungefähren Inhalts
  - z.B. Übersetzungsdienste bei Web-Suchmaschinen

Unterschiedliche Ansätze

- Direkte Übersetzung ohne syntaktische Analyse
  - Gute statistische Systeme, welche aus Paralleltexten lernen
- Übersetzung via Transfer der syntaktischen Struktur (gängige Systeme)
- Übersetzung via semantische Interlingua (theoretisch guter Ansatz)

Skala der maschinellen Übersetzbarkeit

Textsorten geordnet nach Schwierigkeitsgrad der MÜ

1. Wetterberichte, Börsenberichte *
2. Technische Dokumente, Handbücher **
3. Rechtsdokumente **
4. Wissenschaftliche und technische Texte **
5. Journalistische Texte ***
6. Literarische Texte, Werbetexte, Filmtexte ***

Automatisierbarkeit	
*	ohne Nachredaktion
**	mit Nachredaktion
***	zur Zeit unmöglich

Legende

Aber

- Übersetzungsgedächtnisse (translation memory)
  - Auch Kategorie 2/3 ist weitgehend automatisierbar, wenn für eine neuere Version eines Dokuments Übersetzungsteile (Sätze) aus älteren Versionen rezykliert werden.

Textzusammenfassung

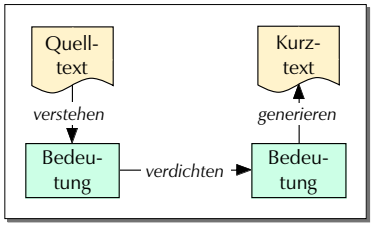
«Language understanding is somewhat like counting from one to infinity; language generation is like counting from infinity to one.» (Y. Wilks)

Echte Textzusammenfassung

- Der Text der Kurzfassung wird erzeugt ausgehend von einer semantischen Form.
- Sehr schwierig! Einfachere Variante…

▸ Informationsextraktion

- Herausfiltern fixer Sachverhaltsmuster
- Beispiel "Übernahmefakten": Welche Firma hat wann welche andere Firma für wieviel Geld übernommen?



Grundmodell des verstehenden Zusammenfassen

Textkompression (TK)

- Aus einem Dokument werden die wichtigsten Sätze extrahiert.

Ansätze für Textzusammenfassung

Statistische und heuristische Verfahren (TK)

- 1958 Luhn: Vorkommen von Wörtern mittlerer Häufigkeit
- 1969 Edmundson:
  - Satzposition im Abschnitt (Anfang/Ende wichtiger als Mitte)
  - wichtige Schlüsselwörter (z.B. aus Überschriften) vs. unwichtige Füll- bzw. Stoppwörter
- 1995 Kupiec: zusätzlich zu Edmundsons Kriterien
  - Satzlänge > 5; Wortmerkmale (Grossschreibung, Länge, Häufigkeit)
  - Textsortenspezifische Indikatorphrasen für zentrale Aussagen "In conclusion,…"
  - Lernendes System, das aus bestehenden Abstract-Dokument-Paaren generalisiert!
    - 80% der Sätze in den Abstracts (von professionellen Zusammenfassern!) waren wörtlich oder nur minim modifiziert im Dokument!

Linguistische und wissensverarbeitende Verfahren

- seit 70er: Meist sehr anwendungsspezifisch!

Telegraphische Verkürzung

Satzbasiertes Auslassen "unwichtiger" Element

- Nur Subjekte, Objekte, Verbalkerne, Präpositionen und abhängige Nominalkerne
- Problem: Keine 1:1-Beziehung zwischen syntaktischer und inhaltlicher Kernfunktion!

- Weiterentwicklungen
  - Weglassen von inhaltlich unwichtigen Satzbestandteilen unter Beibehaltung der syntaktischen Wohlgeformtheit

Text summarization is usually taken to mean **producing** a shorter **version of** an original **document** by **retaining** the most salient **parts of** the original **text**. **Two approaches** have been **favored**: selecting high content-bearing **sentences influenced by** positional **constraints**, **and** performing domain dependent information **extraction** which **fills** a **template from** which a **summary** can be **glossed**.

Text summarization producing version of document by retaining parts of text. Two approaches favored sentences influenced by constraints and extraction fills template from summary glossed.

G. Grefenstette (1998): *Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind*

Fazit Textzusammenfassung

Was sind gute Textzusammenfassungssysteme?

- Relevanz
  - Ist das Wichtige drin? Ist das Unwichtige draussen? Gibt es keine Redundanz?
- Lesbarkeit
  - Textkohärenz: Wie fügt sich das Ganze zusammen? Fehlen Bezugsausdrücke von anaphorischen Ausdrücken? (*dangling pronouns*)

- Seit Mitte 90er-Jahre
  - Verschiedene Evaluations-Konferenzen, wo verschiedene Systeme sich an einer gemeinsamen Aufgabe messen

Zusammenfassung (*Abstract*) vs. Kompression (*Extract*)

- Nach Black(1988) sind gute *Extracts* mühsamer zu lesen – gemäss seinem Experiment zur Verständnisüberprüfungen aber gleichwertig zu *Abstracts*.

Verallgemeinerungsprobleme der CL

Multilingualität

- Viele Techniken funktionieren für eine Sprache wie Englisch.
- Aber sie lassen sich nicht auf andere Sprachen verallgemeinern.

Unbeschränkte Anwendungsgebiete

- Viele Techniken lassen sich in einem eingeschränkten Bereich erfolgreich implementieren.
- Aber sie lassen sich nur beschränkt auf andere oder beliebige Gebiete erweitern.