Einführung in die Computerlinguistik
Überblick

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

2015-10-12

Was ist Computerlinguistik?

Definition

Computational linguistics is the scientific study of models and methods for automatic processing of natural language. Computational linguistics is an interdisciplinary field that shares a large part of its subject matter with computer science and linguistics. However, computational linguists also work on theories, models and methods that are not part of core linguistics or core computer science.

- Closely related, but different: natural language processing, speech recognition

Anwendungen der Computerlinguistik

- Rechtschreibkorrektur
- Grammatiküberprüfung
- Häufigkeitsanalysen von Vorkommen von Wörtern und linguistischen Phänomenen
- Lexikographie (Thesauri, Wörterbücher)
- Informationserschließung (Internet-Recherche)
- Kommunikation mit Maschine z.B. bei der Bank
- (vollautomatische) maschinelle Übersetzung

Beispiele für industrielle Anwendungen

- Internetsuchmaschinen: sehr große Menge an Information, aber hochgradig unstrukturiert $\rightarrow$ direkter Zugang zu relevanten Daten ist schwierig.
- Dialoganwendungen: Zugang zu komplexen Systemen, z.B. Bestellung eines Bahn- oder Flugtickets, Interaktion mit Bank, auch mit natürlichsprachlichen Anwendungen
- Übersetzungssysteme: fremdsprachige Web-Seiten, Gebrauchsanweisungen, Wetterberichte etc.
- automatische Silbentrennung, Rechtschreibprüfung und -korrektur
- automatische Spracherkennung
- Informationsextraktion, z.B. relevante Qualifikationen aus Bewerbungsbriefen und Lebensläufen maschinell extrahieren

Berufsfelder für Computerlinguisten

- Verarbeitung gesprochener Sprache für die Interaktion mit Computern
- Verarbeitung von Texten (suchen, bearbeiten und verwalten)
- Einsatz sprachtechnologischer Software und Ressourcen (in Verlagen, Übersetzungsbüros, Verwaltungen etc.): Maschinelle Übersetzung, elektronische Wörterbücher, Spracherkennung, Sprachgenerierung, lexikonbasierte Optimierung von Optical-Character-Recognition-Verfahren (OCR)
- akademischer Bereich
- Bedarf an Experten steigt tendentiell

Typische Forschungsgegenstände

- Entwicklung von Methoden (Theorie)
- Entwicklung realistischer Anwendungen (Praxis)
- Aufbau und Verwaltung großer wiederverwendbarer Korpora (Daten)
- Konzeption effektiver Evaluationsmechanismen (Experimente)

Nachbardisziplinen (1)

- Linguistik
  - Die Wissenschaft, die sich mit menschlicher Sprache beschäftigt
  - Grundinventar linguistischer Termini
  - Teilgebiete: Phonetik/Phonologie, Morphologie, Syntax, Semantik, Pragmatik; Korpuslinguistik
- Informatik (Algorithmen, Datenstrukturen, Software Engineering)
- Philosophie (Verbindung von Sprache, Denken und Handeln; Relation zu außersprachlichen Gegebenheiten)
- Künstliche Intelligenz (knowledge representation, reasoning, learning)

Nachbardisziplinen (2)

- Kognitionswissenschaft (Sprachbeherrschung ist spezieller Teilbereich der kognitiven Fähigkeiten des Menschen)
- Mathematik
- Insbesondere: Logik, Wahrscheinlichkeitstheorie, Statistik, Graphentheorie
- Sprache ist oft nicht logisch:
  (1) *Ein großer Berg* vs. *Eine große Ameise* $\rightarrow$ Vagheit des Adjektivs (kein Problem für Menschen) $\rightarrow$ Logik modifizieren in CL

  (2) *Vögel fliegen. / Pinguine sind Vögel. / Pinguine fliegen.* $\rightarrow$ scheinbar widersprüchliche Aussagen (Mensch hat wenig Probleme damit)

Teilgebiete der Linguistik

- Phonetik und Phonologie
- Morphologie
- Syntax
- Semantik
- Pragmatik
- Jedes dieser Teilgebiete hat auch eine Entsprechung in der Computerlinguistik.

Phonetik und Phonologie

- artikulatorische Merkmale
- Lautstruktur natürlicher Sprachen
- Spracherkennung: Erkennung und Produktion gesprochener Sprache
- modellieren, welche Segmente ein Wort enthält und wie sich deren Struktur auf die Aussprache auswirkt
- z.B. wenn ein im Prinzip stimmhafter Konsonant am Wortende stimmlos wird ("Auslautverhärtung"):

  (3) *Dieb* /Diep/ vs. *Diebe* /Diebe/

Morphologie

- Bildung und Struktur von Wörtern
- lexikalische Wurzel von einzelnen Wörtern
- Prozesse, verantwortlich für unterschiedliche Erscheinungsformen an der Oberfläche
- Veränderung der Verwendung und Bedeutung des Wortes durch Oberflächenmodifikationen
- z.B. Suffix *-e* als Pluralmarkierung:

  (4) *Dieb*-*e* $\rightarrow$ Dieb-pl $\rightarrow$ “Mehr als ein Dieb”

Syntax

- Strukturbildung von Sätzen
- traditionell am stärksten vertretene Teildisziplin der Computerlinguistik
- Erkennung von Grammatikalität und darauf folgende Bedeutungserschließung
- z.B.

(5) *Der gewitzte Dieb stahl das Geld.* vs. **Der Dieb gewitzte stahl das Geld.*

Semantik

- Bedeutung sprachlicher Einheiten (Wort, Satz etc.)
- z.B.

  (6) *Die Polizei beschlagnahmte das Diebesgut.* vs. *Das Diebesgut beschlagnahmte die Polizei.*

  → gleiche Bedeutung

Pragmatik

- Zweck einer Äußerung in der Welt, z.B.
  Wissen Sie, wie spät es ist?
- Bestimmung des Bezugs von Wörtern: Antezedens eines Pronomens, z.B.:
  Die Katze schnurrt. Sie hat Hunger.
- implizite Annahmen (Präsuppositionen), z.B.:
  “der Präsident von Frankreich wurde nicht in Paris geboren”
  “der Präsident von Norwegen wurde nicht in Oslo geboren”

Korpuslinguistik

- seit Anfang 1980er
- Fortschritte bei Erkennung gesprochener Sprache
- Wortartendisambiguierung (Tagging)
- syntaktische Analyse (Parsing)
- semantische Lesartendisambiguierung (z.B. *Bank 1* vs. *Bank 2*)
- maschinelle Übersetzung

Text corpus

Definition

A corpus (plural corpora) or text corpus is a large and structured set of texts, nowadays usually electronically stored and processed.

- Corpora are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.
- A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus).
- (from Wikipedia)

Kleine Geschichte der Computerlinguistik (1)

- frühe Entwicklung der Computertechnologie (1930er-, 40er-Jahre): numerische Problemstellungen ("Berechnungen", z.B. ballistische Kurven), auch symbolische Verarbeitungsaufgaben (Dechiffrierung verschlüsselter Nachrichtentexte $\rightarrow$ maschinelle Übersetzung (MÜ) als Spezialfall einer Dekodierungsaufgabe)
- frühe Ansätze der MÜ haben gemeinsame Wurzel: stochastische Informationstheorie (Betrachtung des fremdsprachlichen Textes als Ergebnis der Übertragung einer Nachricht über gestörten Kanal $\rightarrow$ Aufgabe: Rekonstruktion des ursprünglichen Nachrichtentextes)
- Statistische Verfahren wurden dann für Jahrzehnte aufgegeben.

Kleine Geschichte der Computerlinguistik (2)

- Aufgabe von statistischen Verfahren weil
- Chomsky die Unzulänglichkeit der statistischen Verfahren der 50er und 60er für Sprachmodellierung nachweist.
- die Leistungsfähigkeit der damaligen Hardware nicht ausreichte (Beschränkungen bevorzugen symbolische Ansätze)

Literatur und Links

- Jurafsky & Martin: Speech and Language Processing. Pearson Prentice Hall. 2008.
- Manning & Schütze: Foundations of Statistical Natural Language Processing. MIT Press. 1999.
- Carstensen et al.: Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg 2010 (3. Auflage)
- elektronische Version beim EasyProxy der Universitätsbibliothek: https://login.easyproxy.ub.uni-muenchen.de/login