LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Ivana Daskalovska

Willkommen zur Tutorium Einführung in die Computerlinguistik

Kontakt:
ivana.bt.mk@gmail.com

Betreff: EICL

Die Folien basieren auf den Vorlesungsfolien von Prof. Dr. Hinrich Schütze. Aufgetretene Fehler liegen ausschließlich in meiner Verantwortung!

- **Was ist Computerlinguistik?**

• Was ist Computerlinguistik?

- **Computerlinguistik** ist die wissenschaftliche Lehre, die Modelle und Methoden für automatische Bearbeitung natürlicher Sprache entwickelt.

- **Nennen Sie die zwei große Teilbereiche der Computerlinguistik**

- **Nennen Sie die zwei große Teilbereiche der Computerlinguistik**

  - **Theoretische Computerlinguistik**: Teildisziplin der Linguistik, die formale berechenbare Modelle natürlicher Sprache entwickelt, implementiert und untersucht

  - **Angewandte Computerlinguistik**: interdisziplinäres Forschungsgebiet (Linguistik, Informatik), das konkrete Algorithmen für die maschinelle Sprachverarbeitung entwickelt (maschinelle Übersetzung, Spracherkennung ...)

- **Wo findet die Computerlinguistik Anwendung?**

- **Wo findet die Computerlinguistik Anwendung?**
  - Häufigkeitsanalysen von Vorkommen von Wörtern
  - Internetsuchmaschinen
  - Lexikographie
  - Übersetzungssysteme
  - Automatische Spracherkennung
  - Informationsextraktion usw.

Nachbardisziplinen der Computerlinguistik?

Nachbardisziplinen der Computerlinguistik?

- Linguistik
- Informatik
- Philosophie
- Künstliche Intelligenz
- Kognitionswissenschaft
- Mathematik

Was ist Linguistik und womit beschäftigt sie sich?

Was ist Linguistik und womit beschäftigt sie sich?

- **Linguistik** (lateinisch *lingua* ‚Sprache', ‚Zunge'), ist die Wissenschaft, die in verschiedenen Herangehensweisen die menschliche Sprache untersucht.
- Sie untersucht die Sprache als System, ihre einzelnen Bestandteile und Einheiten sowie deren Bedeutungen.
- Des Weiteren beschäftigt sich die Sprachwissenschaft mit Entstehung, Herkunft und geschichtlicher Entwicklung von Sprache, mit ihrer vielseitigen Anwendung in der schriftlichen und mündlichen Kommunikation.

- **Nennen Sie die Teilgebiete der Linguistik und beschreiben Sie sie kurz.**

- **Nennen Sie die Teilgebiete der Linguistik und beschreiben Sie diese kurz.**

  - **Phonetik und Phonologie:**
    - Lautstruktur natürlicher Sprachen
    - Artikulatorische Merkmale
    - ***Computerlinguistische Methoden***: Spracherkennung, Sprachsynthese

  - **Morphologie**
    - Bildung und Struktur von Wörtern
    - Prozesse, die für die unterschiedliche Erscheinungsformen an der Oberfläche verantwortlich sind
    - ***Computerlinguistische Methoden***: Wortzerlegung(DEA), Wortartenbestimmung (HMM)

- **Syntax**
  - Strukturbildung von Sätzen
  - ***Computerlinguistische Methoden***: Parsing, computerlesbare Grammatiken (GFGs)

- **Semantik**
  - Bedeutung sprachlicher Einheiten
  - ***Computerlinguistische Methoden***: Wissensdatenbanken, automatische Semantische Analyse

- **Pragmatik**
  - Zweck einer Äußerung in der Welt
  - ***Computerlinguistische Methoden***: Koreferenzresolution, Kontextmodellierung

➢ Korpuslinguistik

- Methode, die auf alle Beschreibungsebenen angewandt werden kann

- Wortartendisambiguierung (Tagging)
- syntaktische Analyse (Parsing)
- semantische Lesartendisambiguierung
- maschinelle Übersetzung

Was ist ein Korpus?

Was ist ein Korpus?

- Strukturierte Sammlung von Texten (heutzutage meist elektronisch gespeichert)

Was versteht man unter Maschinelle Übersetzung?

- **Was versteht man unter Maschinelle Übersetzung?**

  - **Maschinelle Übersetzung** (MÜ oder MT für engl. *machine translation*), auch als **automatische Übersetzung**, bezeichnet die Übersetzung von Texten durch ein Computerprogramm.

- **Was ist Transfer?**

- **Was ist Transfer?**

  - **Transfer**: Direkte Übersetzung von sprachlichen Elementen, ohne Bedeutungsambiguitäten aufzulösen

- **Was ist Interlingua**?

- **Was ist Interlingua**?

  - **Interlingua**: Sprachunabhängige Repräsentation von Bedeutung, in die Sprache überführt werden kann und umgekehrt

- **Wodurch haben Sprachverarbeitungssysteme Schwierigkeiten?**

- **Wodurch haben Sprachverarbeitungssysteme Schwierigkeiten?**

  - **Variabilität**: Die selbe Bedeutung kann durch viele sprachliche Formen ausgedrückt werden.

  - **Ambiguität**: Dieselbe sprachliche Form kann verschiedene Informationen ausdrücken (erst durch den Kontext kann erschlossen werden, was gemeint ist).

- **Welche Typen von Ambiguität gibt es?**

- **Welche Typen von Ambiguität gibt es?**

  - **Phonetische Ambigität (Homophone):**
    Miene - Mine, Meer - mehr, viel - fiel ⇒ Unterschiedliche Wörter haben dieselbe lautliche Form.

  - **Orthographische Ambiguität (Homographen):** übersetzen - über-setzen, umfahren - um-fahren ⇒ Unterschiedliche Wörter werden gleich geschrieben.

  - **Lexikalische Ambiguität (Homonyme):** Maria geht zur Bank. ⇒ Ein Wort hat mehrere verschiedene Bedeutungen.

  - **Morphologische Ambiguität:** Staub-ecken - Stau-becken ⇒ Eine Wortform kann auf unterschiedliche Arten analysiert werden

- **Strukturelle/syntaktische Ambiguität:**
  Peter fuhr seinen Freund sturzbetrunken nach Hause.
  ⇒ Die Grammatikregeln lassen verschiedene Analysen zur Kombination der Satzelemente zu.

- **Kompositionell-semantische Ambiguität bzw. Skopusambiguität**
  Alle Politiker sind nicht korrupt.
  ⇒Quantifikatoren (alle, jeder, zwei) und Negationen können sich auf verschieden große Satzteile beziehen.

- **Pragmatische Ambiguität**:
  Haben Sie eine Uhr?
  ⇒ Der Bezug einer Aussage zum außerlinguistischen Kontext kann auf mehrere Arten hergestellt werden.

- **Was ist ein Wort?**

- **Was ist der Unterschied zwischen Wortform und Lexem?**

- **Was ist der Unterschied zwischen Wortform und Lexem?**

  - **Wortform:**
    flektierte Form eines Wortes, so wie sie im Text oder in geschriebener Sprache vorkommt.
    “sings”, “schönes”

  - **Ein Lexem:**
    eine Klasse lexikalisch äquivalenter Wortformen. Diese Wortformen repräsentieren das Lexem in verschiedenen Umgebungen.
    L1 = {“sing”, “sings”, “singing”, “sang”, “sung”}

- **Was ist der Unterschied zwischen Token und Type?**

- **Was ist der Unterschied zwischen Token und Type?**

  - **Token / Wortvorkommnis:**
    Konkretes Vorkommen eines Wortes

  - **Type** :
    Ein Type bezeichnet eine Klasse von Token …

    - die nicht unterschieden werden ...,
    - die als Kopien wahrgenommen werden ...,
    - die gleich sind

Es war einmal eine alte Geiß, die hatte sieben junge Geißlein, und hatte sie lieb, wie eine Mutter ihre Kinder lieb hat. Eines Tages wollte sie in den Wald gehen und Futter holen, da rief sie alle sieben herbei und sprach: "Liebe Kinder, ich will hinaus in den Wald, seid auf eurer Hut vor dem Wolf, wenn er hereinkommt, so frisst er euch mit Haut und Haar.

- **Welche Bestimmungskriterien werden für den Begriff „Wort“ berücksichtigt?**

- **Orthographisches Kriterium**

  „Wörter sind sprachliche Einheiten, die als Folgen von Buchstaben zwischen Leerzeichen geschrieben werden“

  Problem: Sprachen ohne Buchstabenschrift, weitere Trennzeichen abtrennbare Präfixe bei zusammengesetzten Verben

- **Phonologisches Kriterium**

  „Wörter sind durch eine spezielle einheitliche Akzentstruktur gekennzeichnet, die sich von der entsprechender Wortgruppen/Phrasen unterscheidet.„

  Problem: präzisere Beschreibung der Intonationsmuster nötig

- **Morphologische Kriterien**

  “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  Problem: Im- und Export, hin und her

  “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
  Problem: nicht flektierbare Wörter

- **Morphosyntaktisches Kriterium**

  “Wörter sind die kleinsten sprachlichen Einheiten, die innerhalb des Satzes permutierbar sind.”
  Problem: syntaktische Regeln lassen oft keine Permutation zu
  “das kleine Haus” ⇒ *“das Haus kleine”

- **Semantische Kriterien**

  "[...] kleinste Einheiten des Inhalts oder der Bedeutung."
  "[...] satzfähiges Lautsymbol mit der Eignung, ein Stück Wirklichkeit zu meinen."
  Problem: Funktionswörter: zu, mehrere Wörter für einen Begriff! roter Faden

- **Intuition des Muttersprachlers**

  Wort = durch Muttersprachler intuitiv erkennbare Basiseinheit des Lexikons
  Problem: Intuition nicht immer klar radfahren vs. Rad fahren