Lemmatisierer

Von:

Max Wieck

Jonas Reinhardt

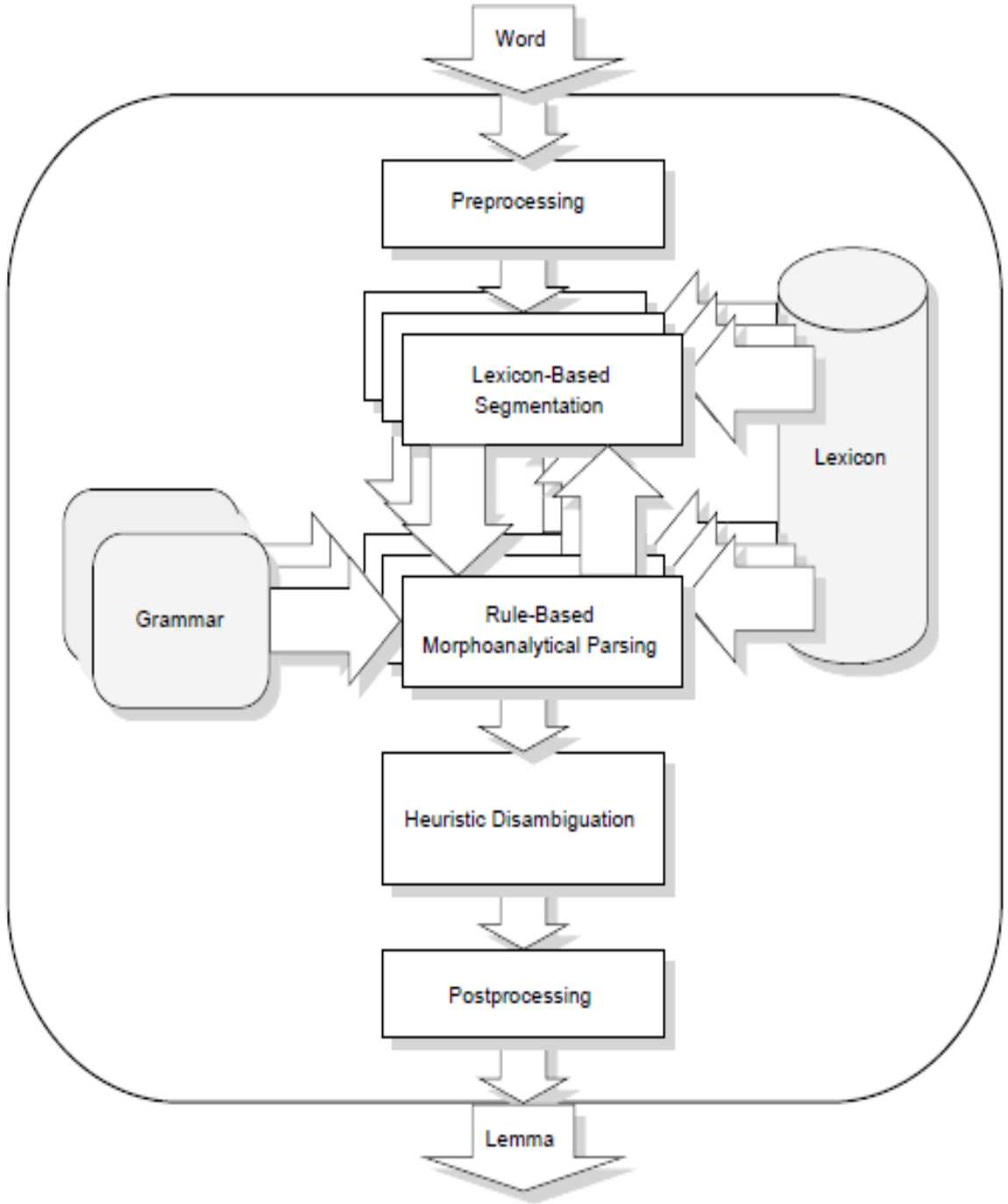
Inhaltsverzeichnis

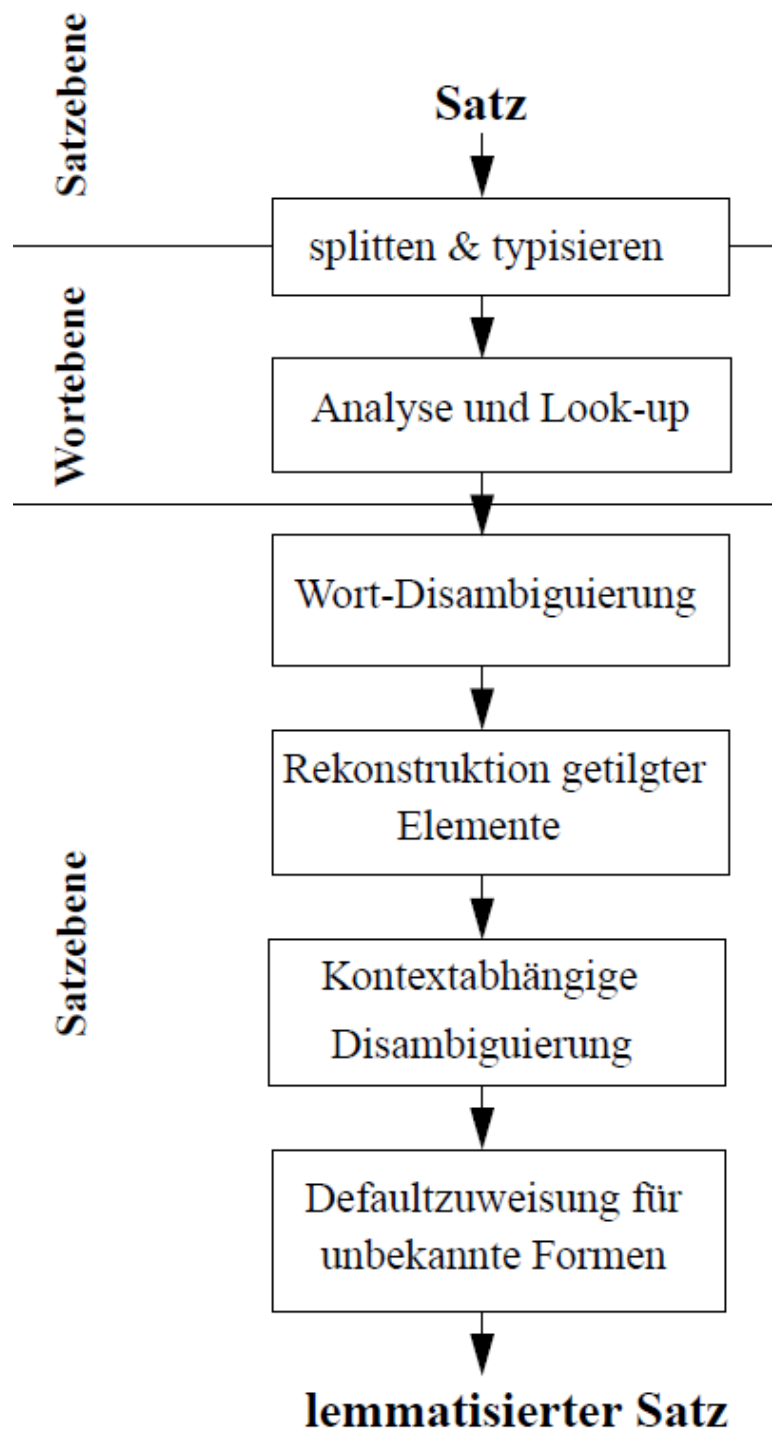
Was sind Lemmatisierer?

- Lemma: Basiswortform/ lexikalische Grundform (Wortform aus dem Lexikon)→ Repräsentant eines Lexems
  - Im Deutschen:
    - Nomen im Nominativ Singular (Bsp.: Traum)
    - Verben im Infinitiv Präsens Aktiv (Bsp.: träumen)
- Lexem: eine sprachliche Bedeutungseinheit
- Lemmatisierer :
  - Klassische/traditionelle Lexikographie: Zuordnung eines Lemmas zu einer Belegform→ untrennbar mit Wörterbuch verknüpft
  - Maschinelle Sprachverarbeitung: (automatische) Rückführung einer Wortform auf die kanonische Grundform zusammen mit einer Annotation linguistischer Informationen

Wofür braucht man Lemmatisierer

- Automatisch lemmatisierte Korpora wichtig für
  - Untersuchungen zu Verbrahmen
  - Verhalten von Komposita
  - Extraktion von Mehrwortlexemen
  - Untersuchungen zum Wortschatz und zur Frequenz bestimmter Lexeme
  - „intelligente“ automatische Rechtschreibkorrektur
- Beispielarbeiten:
  - Strategien der Lemmatisierung von Idiomen
  - Lemmatisierung von Goethes „Wahlverwandschaften“
  - Namenlemmatisierung in der Web-Datenbank mittelalterlicher und frühneuzeitlicher Universitätsmatrikel


Word
Preprocessing
Lexicon-Based Segmentation
Lexicon
Grammar
Rule-Based Morphoanalytical Parsing
Heuristic Disambiguation
Postprocessing
Lemma


Satzebene
Satz
splitten & typisieren
Wortebene
Analyse und Look-up
Wort-Disambiguierung
Rekonstruktion getilgter Elemente
Satzebene
Kontextabhängige Disambiguierung
Defaultzuweisung für unbekannte Formen
lemmatisierter Satz

Wie funktionieren sie? Implementierung

- Preprocessing
  - Preprocessing benötigt um Implementierung verschiedener Eingabetexte zu ermöglichen (→ Darunter fällt: Eliminierung von Wortformen aus dem Fließtext, Silbentrennung verstehen, Normalisierung von Umlauten)
- Lexikon
  - Lemmatisierer extrahieren Informationen aus Wörterbüchern
  - Lexem im Lexikon mit einem/mehrere Einträgen mit alternativen morpho-syntaktischen Paradigmen (Bsp. *Er* als Pronomen, oder als Präfix mit Verben…)
- Segmentierung
  - Segmentierung der Eingabewortformen durch lexikongesteuerte Bottom-Up-Algorithmus für die Morphem-Erkennung
  - Algorithmus erstellt vollständigen Segmentierungsbaum über Rückverfolgung für jedes Eingabewort

- Grammatik und Parser
  - Morpho-analytischer Parser verwendet, um falsch/schlecht gebildete Segmentirungs“lesungen“ zu streichen
  - Parser gewichtet Segmentierungen über Segmentierungsbaum und Ähnlichkeit (Affinitätsinformationen)
- Context Switching
  - Kontextwechseltreibe (context switching driver) steuert Segmentierungs- und Parsingprozess
  - Treiber unterbricht Segmentierungsalgorithmus nachhinzufügen eines neuen Astes im Segmentierungsbaum
  - Morphoanalytischer Parser prüft auf Konformität über Grammatik und schick Info zurück
  - Treiber entscheidet ob neuer Ast schlecht und erneut zurückverfolgt werden muss oder fährt mit Segmentierungsprozess fort
  - Erfolgreich geparste Segmentierungsbäume werden gewichtet und folgend auf Mehrdeutigkeit geprüft

- Disambiguation (Mehrdeutigkeit)
  - Lesarten der Segmentierungsbäume in Wahrscheinlichkeiten eingeordnet und bewertet
    - Lesarten die schlechter als die am besten bewerteten sind, werden rausgeworfem
    - Rangfolge über heuristische Quantifikation des „morphologischen Abstands“ zwischen Eingabewort und Struktur des Segmentierungsbaum
    - Abschließende Extraktion der Lemmas aus den gesäuberten Segmentierungsbäume
- Postprocessing
  - Morpho-analytische Informationen des Eingabewortes auf folgende Strukutr reduziert

<word_form>

<composition_tag><derivation_tag><lemma_name>

Unterscheidung nach verwendetem Lexika

- Vollformenlexikon
  - Alle möglichen Formen eines Lexems werden aufgeführt. → Keine morphologische Analyse beim Look-up einer Wortform nötig.

- Grundformenlexikon
  - Eine zusätzliche morphologische Analyse notwendig, da nicht ohne weiteres ein Programm Wortformen auf ihre Grundformen reduzieren kann.
  - Grundlage eines Vollformenlexikons

Disambiguierungsmethoden

→ Kontextabhängig

- Wortebezogene Lemmatisierung
  - Da sich hier der Kontext nur auf das einzelne Wort beschränkt, fällt die Disambiguierung schwer.
  - Nur Groß- und Kleinschreibung bieten Anhaltspunkte
  - Auch die statisch-basierte Disambiguierung ist eine Option
- Satzbezogene Lemmatisierung
  - Ist der Analyse auf Wortebene weit überlegen.
  - Die Analysetiefe schwankt je nach gewähltem System

Praktisches Anwendungsbeispiel

- https://cst.dk/tools/

Literaturverzeichnis

- Carstensen, Kai-Uwe, Ebert, Christian (Hrsg.): Computerlinguistik und Sprachtechnologie. Eine Einführung. 3. Aufl. Heidelberg 2010.
- Cyril, Belica: WP2-Lemmatizer, Final Report. MLAP93-21 MECOLB. Luxembourg 1994.
- Glück, Helmut, Rödel, Michael (Hrsg.): Metzler Lexikon Sprache. 5.Aufl. Stuttgart 2016.
- https://www.cis.uni-muenchen.de/download/cis-berichte/95-084.pdf