Lemmatisierung

•••

Das Schweigen des Lemma

Ein Vortrag von Tim Eron, Henrik Schütz, Jan Springer

Was ist das?

Bestimmung der Grundform eines Wortes

- **Problem: Wörter können z.B. konjugiert oder dekliniert sein**
- **eine Lösung: *Stemming* - man reduziert Wörter auf den Wortstamm, indem man Affixe entfernt**

Beispiel: “treffend” & “treffen” werden zu “treff”

Problem: Flexionsform “getroffen” würde nicht erkannt werden

Was ist das?

Das Problem haben wir bei der Lemmatisierung nicht

- **Lemma: Grundform eines Wortes innerhalb eines Lexikons (z.B. “ging” wird “gehen” zugeordnet)**
- **Lexikon: abrufbarer Wortbestand einer Sprache**
- **Beim Lemmatisieren wird das Wort auf diese Grundform zurückgeführt und zugeordnet**

Beispiel: “treffend” & “getroffen” werden “treffen” zugeordnet

Wie funktioniert das?

1. **Jedes Wort des Textes wird mit dem Wörterbuch abgeglichen**
2. **Kann es nicht zugeordnet werden, wird geprüft in welche Teile das Wort zerlegt werden kann**
3. **Diese Teilfolgen werden dann wieder mit dem Wörterbuch abgeglichen**

Das passiert so lange bis eine Übereinstimmung gefunden wurde

- Die Zerlegungsprogramme müssen hierarchisiert werden und auf das Wörterbuch angepasst sein, um mögliche Kombinationen festzulegen

Klein, W., & Rath, R. (1981). Automatische Lemmatisierung deutscher Flexionsformen. In *Computer in der Übersetzungswissenschaft* (pp. 94-142). Verlag Peter Lang.

Wie funktioniert das?

gibt es keine Übereinstimmung, soll das Lemma neu im Wörterbuch aufgenommen werden:

1. **Wort wird auf bekannte Folgen analysiert (z.B. -keit, -lich)**
    a. dafür müssen diese dem Programm bekannt sein
2. **Enthält das Wort eine bekannte Folge, kann es lemmatisiert werden**
3. **Enthält es keine bekannte Folge, wird es zu einem künstlichen Homographen erklärt**

Um das Wort weiter zu analysieren könnte man die Umgebung des Wortes analysieren

Klein, W., & Rath, R. (1981). Automatische Lemmatisierung deutscher Flexionsformen. In *Computer in der Übersetzungswissenschaft* (pp. 94-142). Verlag Peter Lang.

Wofür braucht man das?

- Kategorisierung von Texten
- Suchmaschinen, ähnliche Wörter zu einem gegebenem Lemma finden
- Kookkurrenzanalyse
- Diachrone Sprachbetrachtung

Praktische Vorführung eines Systems

Cosmas II

Probleme beim *Stemming*:

Over-Stemming:

- Vom Wort wird zu viel entfernt – dadurch können unterschiedliche Wörter fälschlicherweise unter einen Wortstamm fallen

Under-Stemming:

- Vom Wort wird zu wenig entfernt – zusammengehörende Wörter werden nun unterschiedlichen Wortstämmen zugeordnet

Das Dilemma ...

... beim Wort-Lexikon:

- Längere Verarbeitungszeit / hohe Speicherkapazität
- Wort-Lexikon muss händisch kuratiert werden (bei unregelmäßigen Formen) – *Vorteil* : Ergebnisse sind besser, da sie wahrscheinlich von einem Linguisten überprüft worden sind

Das Dilemma ...

... beim Lemmatisieren:

- Eigennamen werden nicht erkannt (bzw. müssten zusätzlich im Lexikon aufgenommen werden)
- Kompositazerlegung: bei der Trennung kann die semantische Bedeutung verloren gehen Beispiel: All+gemein+wissen; Brief+kasten+schlüssel; Gemeinde+grund+steuer
- Manche Zuordnungen können nur mit einer syntaktischen Analyse erfolgen (Beispiel: “billige”)

Praktische Vorführung eines Systems

Python nltk

Quellen:

- Klein, Wolfgang und Rainer Rath (1981). Automatische Lemmatisierung deutscher Flexionsformen. In *Computer in der Übersetzungswissenschaft* (S. 94-142). Verlag Peter Lang. // https://pure.mpg.de/rest/items/item_468673_3/component/file_468719/content
- Bußmann, H. (1983). Lexikon der Sprachwissenschaft Stuttgart.
- Verschiedene Wort-Lexika (.txt-Dateien): https://github.com/michmech/lemmatization-lists

Over- and Under-Stemming:

- Patil, Harshali B. und Ajay S. Patil (2019). A Hybrid Stemmer for the Affix Stacking Language: Marathi. In *Computing in Engineering and Technology: Proceedings of ICCET 2019* (S. 448). Springer Verlag.
- https://pdfs.semanticscholar.org/1c0c/0fa35d4ff8a2f925eb955e48d655494bd167.pdf (4. Errors in Stemming)

Quellen:

- search.aol.com & de.spongepedia.org
- Allgemeinwissen
- hab ich gehört
- Skandinavische Wissenschaftler (z.B. Prof. Eide)
- Probs an Mr. Dr. Andy Witt
- Wir haben die PoS-Tagger Gruppe abgehört, die konnten das halt und wir wollten nicht selber googlen
- Ted von TedTalk