

Verschlagwortung mit Wikipedia

5. Semantics Day
am Fraunhofer-Institut für Graphische Datenverarbeitung

Wikipedia als Thesaurus für sprachunabhängige Verschlagwortung

Daniel Kinzler
Wikimedia Deutschland



Verschlagwortung

- Vokabular von Schlagwörtern
- "Vokabular" unabhängig von Sprache: abstrakte Bezeichner für Konzepte (IDs, URIs)
- *Bezeichnungen* für jedes Konzept (in verschiedenen Sprachen)
- Insbesondere auch Spezialbegriffe

Wikipedia

- Wikipedia: mehrere Millionen Konzepte
- Anders als Wörterbuch:
 - Orte, Personen, Lebewesen, Organisationen, Produkte, Spezialbegriffe
- Wikipedia: über hundert Sprachen, untereinander verknüpft

WikiWord

- Multilingualer Thesaurus aus Wikipedia
- *Konzeptbasierter* Thesaurus:
 - **Beziehungen** zwischen Konzepten
 - **Bezeichnungen** für Konzepte (verschiedene Sprachen)
 - **Definitionen** von Konzepten (verschiedene Sprachen)
- Diplomarbeit *"Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia"* (Uni Leipzig, 2008)

WikiWord

- **Kategorien:** Über- bzw. Unterordnung von Konzepten
- **Querverweise:** Verwandtheit von Konzepten
- **Erster Satz:** Definition
- Seitentitel, Weiterleitungen und vor allem **Link-Text** ("Ankertext"): Bezeichnungen

`[[Primärschlüssel | Key]]`

WikiWord

- Grobe Klassifikation:
 - Person, Ort, Lebensform, Zeitpunkt, etc.
- Schlüsseleigenschaften aus Templates ("Infoboxen"), z.B.:
 - Geburtsdatum und PND von Personen
 - Koordinaten von Orten
 - CAS-Nummer für Chemikalien, etc

WikiWord

- Zunächst ein Thesaurus pro Sprache
- Vereinigung: Beschreibungen des selben Konzeptes in verschiedenen Sprachen zusammenfassen
- "Inter-Language-Links" verbinden Wikipedia-Artikel in unterschiedlichen Sprachen

WikiWord

- Für Englisch, Deutsch, Französisch, Niederländisch, Norwegisch:
 - >20 millionen Terme
 - >11 millionen Konzepte
 - >2 millionen Definitionen
 - >75 millionen Querverweise
 - >11 millionen Unterordnungen
- Export als SKOS

Fragen an den Thesaurus

- *Welche verschiedenen Bedeutungen gibt es für einen Ausdruck?*
- *Welche Bezeichnungen gibt es für ein bestimmtes Ding bzw. Konzept?*
- *Welche Konzepte sind einem gegebenen Konzept inhaltlich verwandt?*
- *Wie nahe stehen sich zwei Konzepte inhaltlich?*
- *Welche Bedeutung ist in einem gegebenen Kontext die wahrscheinlichste?*

Qualität des Thesaurus

- Bezeichnungen:
 - Ohne Filter:
 - ca. 93% der Terme korrekt
 - ca. 97% der Verwendungen korrekt
 - Mit Filter:
 - ca. 94% der Terme korrekt (80% Abdeckung)
 - ca. 98% der Verwendungen korrekt (95% Abdeckung)
- Semantische Nähe:
 - Korrelationswert von 0,41 (r-Wert nach Person)

Disambiguierung

Maximierung der Kohärenz

- 1) Alle möglichen Bedeutungen für jeden Ausdruck
- 2) Alle möglichen Kombinationen von Bedeutungen (*Interpretationen*)
- 3) Für jede Kombination: Summe der semantischen Nähe (*proximity*) aller Paare von Bedeutungen (*Kohärenzwert*)
- 4) Wähle Interpretation mit der größten inneren Nähe (*Kohärenz*)

Disambiguierung

"Mond und Erde"

- 1) "Mond": Erdmond, Satellit, Mond (Heraldik), ...
"Erde": Planet Erde, Erdung, Boden, Humus, ...
- 2) (Erdmond/Planet Erde), (Erdmond/Erdung),
(Satellit/Boden), ...
- 3) (Erdmond/Planet Erde) = 0.9,
(Erdmond/Erdung) = 0.2, (Satellit/Boden) = 0
- 4) "Mond" = Erdmond, "Erde" = Planet Erde

Wikimedia Commons

- Archiv für freie Medien (vor allem Bilder)
- Beschreibungen und Verschlagwortung größtenteils auf Englisch
- Gewünscht: Suche in anderen Sprachen

Wikimedia Commons

- Schlagwörter (Kategorien) auf Commons entsprechen Wikipedia-Kategorien
- Direkte Zuordnung von WikiWord-Konzepten zu Kategorien auf Commons
- Suchergebnis: alle möglichen Bedeutungen, jeweils Definition und Beispiel-Bilder

Bridgeman Art Library

- Bildagentur, spezialisiert auf Kunst
- Großes Archiv, Beschreibung und Verschlagwortung manuell, auf Englisch
- Gewünscht:
 - Systematisierung der Verschlagwortung
 - Synonyme und Übersetzungen für die Suche
- *Kein Wikimedia-Projekt*

Bridgeman Art Library

- Informationen kann WikiWord liefern
- Aufgabe: Schlagwörtern die *richtigen* Bedeutungen zuordnen
 - Mehrdeutigkeiten auflösen anhand des Kontextes
 - Wissen über Eigenschaften nutzen, z.B. Entstehungsdatum, Lebensdaten von Künstlern

Bundesarchiv und Fotothek

- 100.000 Bilder vom Bundesarchiv, 250.000 von der Deutschen Fotothek
- Metadaten: Titel, Schlagwörter, Zeit, Ort, etc
- WikiWord: findet Kategorien auf Commons
- Aufgabe: Disambiguierung der Schlagwörter

WikiWord

- WikiWord wird laufend weiterentwickelt
- Alle Anwendungen noch in der Entwicklungsphase
- Systematische Evaluation fehlt noch

Verschlagwortung mit Wikipedia

Vielen Dank!

**Wikimedia Deutschland
Gesellschaft zur Förderung Freien Wissens e. V.**

<http://wikimedia.de>

<http://brightbyte.de/page/WikiWord>

