Parallele Korpora und Terminologie-Extraktion

Martin Volk
Universität Zürich

Überblick

1. Terminologie-Datenbanken
2. Terminologie-Extraktion
3. Alignierung
4. Translation-Memory (Übersetzungsarchiv)

Was ist Terminologie?

- Terminologie ist die systematische Untersuchung technischer Terme. Dazu gehört:
  - die Sammlung, Beschreibung, Verarbeitung und Präsentation von Termen, die eine besondere Verwendung erfahren, in einer oder mehreren Sprachen.
- Vorsicht: Mehrdeutigkeit!
  - Terminologie = Wissenschaft
  - Terminologie = Sammlung von Termen

Was ist ein Term?

- Annäherung: Alle Wörter, die nicht in einem allgemeinsprachlichen Wörterbuch verzeichnet sind (z.B. *router*, *romvelope*).
- Genauer: Alle Wörter (Bezeichnungen), die nicht in einem allgemeinsprachlichen Wörterbuch verzeichnet sind und/oder in einer fachspezifischen Bedeutung benutzt werden (z.B. *Zucker* in Haushalt/Chemie).

Terminologie-Datenbank

ist eine Datenbank für Fachbegriffe mit u.a. den Datenfeldern:

- Benennung (in verschiedenen Sprachen), Orthographische Variante
- Fachgebiet
- Definition, Kontextbeispiele, Querverweise
- Quellenangaben
- Synonyme, Antonyme, Hyponyme → Thesaurus-Relationen
- Symbole, Graphiken, Bilder
- Grammatische Informationen (Genus,
- Verwaltungsinfo: Datum, Autor, Status, ...

- für eine umfassende Liste siehe: www.ttt.org/clsframe (ISO 12620)

Terminologie der Terminologie

- **Begriffe** als Denkeinheiten werden sprachlich repräsentiert durch **Bezeichnungen**.
- Als **Bezeichnungen** gelten (nach F. Mayer S.65)
  - Benennungen
    - Einwortbenennung
    - Mehrwortbenennung
  - Symbole (z.B. © / ®)
  - Nummern (z.B. 360° / 37°)
  - Notationen (z.B. H^2O)

Terminologie-Datenbanken

Beispiele für Terminologie-Datenbanken:

- Eurodicautom (die Terminologie-DB der EU; http://europa.eu.int/eurodicautom/login.jsp)
- Medical terms in 9 European Languages (http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html)

Beispiele für Terminologie-Datenbanksysteme:

- MultiTerm (Trados)
- TermStar (STAR)
- TermOrganizer (in der XTS, Xerox)

Terminologie-Datenbanken

- Vorteile:
  - schneller Zugriff
  - über Netzwerk für Teamarbeit nutzbar
  - Konsistenzsicherung über Zeit und Mitarbeiter
- Nachteil:
  - Datenpflege ist aufwendig!

Terminologie-Extraktion

= die automatische Gewinnung von Term-Kandidaten aus einem Text (kurz: Termextraktion).

- Erste Systeme auf dem Markt:
  - Xerox TermFinder
  - System Quirk
  - Trados ExtraTerm
- Problem: Systeme sind noch ungenau.
- Studie der Uni Zürich vom Dez. 1999:
  - Präzision: Nur 20% aller gelieferten Termkandidaten sind Terme.
  - Vollständigkeit: Nur 80% aller Terme eines Textes werden gefunden.

Terminologie-Extraktion

ist verwandt mit

- der automatischen Erstellung eines Indexes am Ende eines Buches (*back of the book index*)
- der automatischen Verschlagwortung von Büchern (*keyword assignment*)
- der automatischen Indexierung für Suchmaschinen (Information Retrieval)

Methoden der Termextraktion

- linguistische Methoden
  1. Bestimmung charakteristischer Morpheme und aller damit gebildeter Komposita.

  Beispiel aus der Elektrotechnik:

  Impuls →
  - *Zählimpuls, Rechteckimpuls*
  - *Hardwareimpuls*
  - *Steuerimpuls*
  - *Impulsbreite, Impulshöhe, Impulszeit*
  - *Impulsdauer, Impulsverhalten*
  - *24-V-Impulsgeber*

Methoden der Termextraktion

- linguistische Methoden
  2. Bestimmung der Wortarten und der sog. Nominalphrasen (siehe TermFinder)
- statistische Methoden (meist kombiniert mit Wortartenbestimmung)
  - Bestimmung der Häufigkeit aller Substantive und aller Adjektiv-Substantiv-Paare in einem Dokument und in allen (?) Dokumenten.
  - Terme kommen im Dokument häufig und in der Gesamtmenge der Dokumente selten vor.

Schritte bei der Termextraktion

1. Bestimme die Term-Kandidaten
2. Filtere die Term-Kandidaten. Aussortieren von
   - Funktionsverbgefügen (*in Rechnung stellen*)
   - temporalen Ausdrücken (*Ende des Monats*)
   - idiomatischen Wendungen (*ein Wink mit dem Zaunpfahl*)
3. Extrahiere Teile von komplexen Termen als potentielle Terme
4. Bestimme Beziehungen zwischen Termen (Oberbegriff - Unterbegriff; Teil - Ganzes)

Masszahlen

beschreiben die Qualität der Termextraktion

- Vollständigkeit (*Recall*)
  - Ziel:
    - möglichst alle Terme finden
    - möglichst wenige Terme übersehen
- Genauigkeit (*Precision*)
  - Ziel:
    - möglichst nur wirkliche Fachterme finden
    - möglichst wenige allgemeinsprachliche Wörter als Term klassifizieren

Arbeitsweise des Xerox TermFinders

Bei **monolingualer** Term-Extraktion

1. Der Text wird in Worte und Satzzeichen (genauer: Zeichenketten) zerhackt (einschliesslich der Unterscheidung Satzendepunkte vs. Abkürzungspunkte und Ordinalzahlenpunkte).
2. Jedes Wort wird morphologisch analysiert. (Basis: ein grosses Lexikon mit Wortstämmen): Das Programm liefert alle möglichen Lesarten und die zugehörigen Grundformen zu jeder Wortform. Wörter, die nicht im Lexikon sind, werden geraten.
3. Die Lesarten werden mit Hilfe des Kontextes disambiguiert. (= *Part-of-Speech Tagging;* erreicht ca. 95% Genauigkeit)
4. Alle Nominalphrasen (NP = Nomen mit seinen Attributen) werden extrahiert.

Arbeitsweise des Xerox TermFinders

- TermFinder findet folgende Arten von Nominalphrasen in englischen Texten
  - noun: *account*
  - adjective + noun: *non-financial enterprise*
  - compound noun: *interbank market*
  - adjective + compound noun:
    - *EMU-wide interbank market*
    - *uniform money market interest rate*
  - noun + *of* - prepositional phrase:
    - *settlement of cross-border payments*

Arbeitsweise des Xerox TermFinders

Bei **bilingualer** Term-Extraktion

1. Die Texte werden aligniert (d.h. es werden parallele Übersetzungseinheiten gebildet).
2. Bei beiden Texten werden die Term-Kandidaten extrahiert wie bei monolingualer Term-Extraktion.
3. Jeder Term-Kandidat aus der A-Text-Einheit bildet mit jedem Term-Kandidaten aus der parallelen Z-Text-Einheit ein Term-Kandidaten-Paar.

Vorteile des Xerox TermFinders

- Berechnet Grundform eines Wortes
- Viele Sprachen (DE, EN, FR, IT, ES, PT, NL)
- Kombination mit der DB TermOrganizer
  - Term-Kandidaten in DB
  - Grundform in DB
  - Auftretenshäufigkeit in DB
  - Kontext in DB
  - Bei Mehrwortausdruck: *Headword* in DB

Probleme des Xerox TermFinders

- Bestehende Term-Sammlungen werden bei der Extraktion nicht berücksichtigt.
- Die Häufigkeit eines Term-Kandidaten bzgl. eines allgemeinsprachlichen Korpus wird nicht berücksichtigt.
- Die interne Struktur eines Term-Kandidaten wird nicht berücksichtigt.
- Bei bilingualer Extraktion:
  - Die Strukturäquivalenz wird nicht berücksichtigt.
  - Die Länge eines Term-Kandidaten im Vergleich zu der Länge der Term-Entsprechung wird nicht berücksichtigt.

Datenbank

- In einer Datenbank (DB) werden (grosse Mengen) Daten strukturiert abgelegt.
- In einer Terminologie-Datenbank wird Fachterminologie abgelegt.
  - Benennungen (in mehreren Sprachen)
  - Definitionen
  - Quellenangaben
  - grammatische Angaben
- In einer **Translation-Memory Datenbank** werden Übersetzungseinheiten (typischerweise bilinguale Satzpaare) abgelegt.

Translation Memory - Systeme

- auch: Übersetzungsspeicher, -datenbank oder -archiv
- kurz: TM-System

Beispiele: Trados Workbench, STAR Transit, Déjà Vu, SDLX, IBM Translation Manager

Idee: Übersetzte Sätze werden paarweise abgelegt und bei weiteren Übersetzungen genutzt.

Wichtig: Auch ähnliche Sätze werden im TM gefunden!

TM-Systeme (Beispiel)

Im Translation Memory	
The generated pulses are converted into on-off signals which are sent to the ECU to control fuel injection.	Les impulsions produites sont converties en signaux de marche/arrêt qui sont a leur tour transmis à l'ECU pour commander l'injection du carburant.
Neu zu übersetzen	
The generated pulses are converted into on-off signals which are sent to the ECU to control fuel injection and ignition timing.	??

Auffüllen eines TM-Systems

- Bei der manuellen Übersetzung: Satz für Satz in die DB ablegen.
- Automatisches Laden bestehender Übersetzungen (Ausgangstext - Zieltext)
  - Problem: Wie findet man zusammengehörende Übersetzungseinheiten?
  - Lösung: Alignierungsprogramme
    - parallelisieren bestehende Übersetzungen und bilden bilinguale Übersetzungseinheiten.
    - basieren auf: Vergleich von Textstruktur (Absätze, Überschriften etc.) und Satzlängen

Alignierungsprogramme

- z.B. WinAlign (Fa. Trados), sonst häufig in Verbindung mit Translation-Memory Produkten oder auch bil. Term-Extraktion
- müssen viele Dateiformate verarbeiten können (Word, HTML, Framemaker, Quarkexpress, PageMaker)
- müssen schnelles, manuelles Nachbearbeiten unterstützen

Struktur eines TM-Systems

- Unterscheide: uni- bzw. bidirektionale TM-Systeme
- 1:1 vs. 1:n TM-Systeme
- Vergabe von Zugriffsrechten
- Automatische Zuweisung von Systeminformationen (Datum, Benutzer)
- Manuelle Strukturierung der DB durch Attribute (Sachgebiet, Quelle, Status etc.)

Abfragen eines TM-Systems

- Manuelles Nachschlagen (sog. Konkordanz-Information)
  - von Sätzen und
  - von Einzelwörtern (in allen morphologischen Formen)
- Automatisches Nachschlagen von (ähnlichen) Sätzen (Fuzzy Match = unscharfer Vergleich)
  - Anzeigen des Ähnlichkeitsgrades
  - Anzeigen der Unterschiede zwischen gesuchtem und gefundenem Satz
  - Vorsicht: Berechnung der Ähnlichkeit beruht auf statistischen Verfahren (nicht auf linguistischen)

Änderungsoperationen

- Hinzufügen
- Weglassen
- Umstellung (Änderung der Reihenfolge)
- Ersetzen (an gegebener Position)

- Mechanische Operationen
- Linguistische Operationen

Einflüsse auf den Fuzzy-Match (bei der Trados Workbench)

- Anzahl der geänderten Wörter
- Umstellung (auch bei gleichbleibender Bedeutung)
  - Aus dem Menü **Datei** wählen Sie den Befehl **NEU.**
  - Wählen Sie aus dem Menü **Datei** den Befehl **NEU.** → 74% Match
  - Nötiges NLP-Modul: NP/PP-Chunker

Einflüsse auf den Fuzzy-Match (bei der Trados Workbench)

- Satzlänge (Anzahl der Wörter)
- Satzanfang wird besonders gewichtet.
- Formatierung (geringer Einfluss)
- Satzzeichen (geringer Einfluss)

Keine Einflüsse auf den Fuzzy-Match (bei der Trados Workbench)

- Länge der geänderten Wörter
- Typ der geänderten Wörter (Adjektiv, Substantiv, etc.)
  - TM: Wählen Sie aus dem Menü **Datei** den Befehl **NEU.**
  - Klicken Sie im Menü **Datei** auf den Befehl **NEU.** → 57%
  - Starten Sie den Befehl **NEU** über das Menü **Datei.** → 40%
  - Nötiges NLP-Modul: PoS-Tagger

Keine Einflüsse auf den Fuzzy-Match (bei der Trados Workbench)

- Flexionsvarianten
  - TM: `Das ausgewählte Segment oder Feld darf nur exakt diesen Text enthalten.`
  - `Die ausgewählten Segmente oder Felder dürfen nur exakt diesen Text enthalten.` → 73%
  - Nötiges NLP-Modul: Morphologie-Analyse (Lemmatisierung)

Keine Einflüsse auf den Fuzzy-Match (bei der Trados Workbench)

- Passivierung
- Satzstruktur
  - `TM: Beim Laden eines Stapels von Seiten im Hochformat müssen die bedruckten Seiten nach unten zeigen, und der Kopf der Seite muss zuerst in den Scanner eingeführt werden.`
  - `Beim Laden eines Stapels von Seiten im Hochformat müssen die bedruckten Seiten nach unten zeigen.`
  - Nötiges NLP-Modul: Clause-Grenzen-Erkenner

Translation Memory-Systeme

- Vorteile:
  - Vermeidung von Arbeitswiederholungen
  - Zeitersparnis bei repetitiven Texten oder neuen Textversionen
  - Konsistenzsicherung über Zeit und Mitarbeiter
- Probleme:
  - Gefahr für Brüche in der Textkohärenz
  - Datenpflege: Qualität der Einträge

Einordnung

- **Terminologie-DB**: speichert und findet Einzelwörter und Ausdrücke
- **Translation Memory-System**: speichert und findet ganze Sätze
- z.Zt. **fehlend**: DB für Satzteile (z.B. Relativsätze oder andere Nebensätze)
  - Probleme:
    - Auffüllen: automatisches Erkennen der Satzteile
    - Anwenden: Erkennen der Satzteile im A-Text und Zusammensetzen im Z-Text