Automatische Termextraktion (TE): »Utopie, Kommerz, Ansätze, Zukunft«

Simon Clematide
Institut für Computerlinguistik
Universität Zürich

13.3.2004
http://www.cl.unizh.ch/siclemat/talks/termext04/

Das Programm

Übersicht

- Utopien vollautomatischer TE
- TE als Identifikation von Term-Kandidaten
  - Probleme der vollautomatischen TE
- Von Zeichenketten zu Konzepten
  - Segmentieren und Klassifizieren
- Vergleichende Evaluation von 3 kommerziellen TE-Tools
  - Silence und Noise
- Methoden der TE
  - Linguistisch und quantitativ/statistisch
- Nähere Zukunft
  - Anwendungen und Methoden

Die monolinguale Utopie

Vollautomatisches Erarbeiten aller relevanten Fachwörter aus einsprachigen, elektronisch gespeicherten Texten

- per Knopfdruck
- mit höchster Genauigkeit
- unabhängig von Sprache, Fachgebiet, Datei- und Textformat
- mit Einbezug bereits vorhandener Terminologiebestände
- inklusive terminologischer Varianten
- inklusive wichtiger linguistischer Merkmale wie Grundform, Wortart bzw. Wörterstruktur, Verwendungsrestriktionen, ...
- inklusive Häufigkeiten, typischen Belegstellen, Quellenangabe
- unter optimaler Einbettung in den Arbeitsablauf (*work flow*)

Die bilinguale Utopie

Vollautomatisches Erarbeiten der Übersetzungspaare aller relevanten Fachwörter aus zweisprachigen Parallel-Texten

- mit der Qualität der monolingualen Termextraktion
- inklusive Zuordnung von linguistischen Merkmalen über die Sprachen hinweg
- mit Übersetzungsvarianten
- inklusive Angaben zu Häufigkeit und Belegstellen der Übersetzungspaare

Termextraktion nüchtern betrachtet

Automatische Termextraktion
=
»Computergestütztes Identifizieren von potentiellen terminologischen Einträgen (Termkandidaten)«

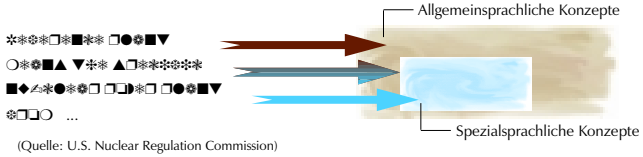
Warum so bescheiden?

- technisches Problem: unzureichende Sprachtechnologie
- linguistisches Problem: mangelhafte Kriterien für Termhaftigkeit
- philosophisches Problem: Was sind *relevante* Einträge eines Fach- bzw. Spezialvokabulars?
  - Abhängig von Verwendungszweck (normativer bzw. deskriptiver Terminologie-aufbau, Übersetzung, Informationssuchsysteme) und vom intendierten Zielpublikum

Von Zeichenketten zu Konzepten

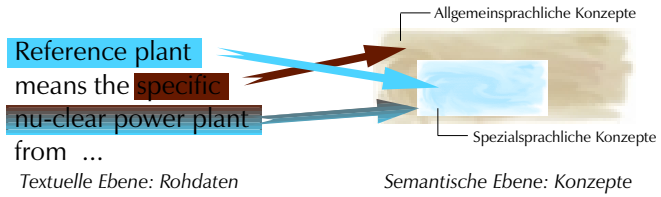
Termini à la ISO: »special language concept designator«

- Termini konstituieren sich über ihre semantische Funktion
- Knacknuss: Wie kann der Computer als »Nicht-Muttersprachler« rohe elektronischen Textdaten auf spezialsprachliche Bedeutungen beziehen?



- Welche Einheiten der Texte verweisen auf welche Art von Konzepten?

Textteile segmentieren & klassifizieren



- Segmentieren
  - Wo beginnen und wo enden ein- oder mehrwortige Textsegmente?
    - Identifikation von Subsegmenten: »nuclear power«, »power plant«
- Klassifizieren
  - Welche Textsegmente beziehen sich auf spezialsprachliche Konzepte?
    - oft eine graduelle und textsortenabhängige Einordnung

Segment- und Klassifikationsfehler

Mangelhafte Segmentierung (*noise*)

- Termkandidat enthält teilweise nicht-terminologisches Material
  »relevant normname«
- Termkandidat enthält mehrere Termini
  »normname and type«
- Termkandidat enthält Terminus nicht vollständig
  von »WORK mode« wurde nur der Anfang in »return to WORK« gefunden

Falsche Klassifikation

- Termkandidat ist kein Terminus (*noise*)
  »according«
- vorhandener Terminus erzeugt keinen Termkandidaten (*silence*)
  »status messages«

Monolinguales TE-Evaluationsbeispiel

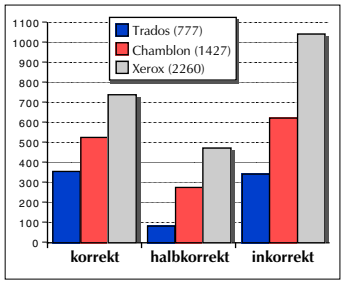
Vergleichende Studie von 3 kommerziellen TE-Tools

- Anwendungsbereich: Technische Redaktion/Software
- Fachtexte: 2 englische Handbücher (total ca. 200 Seiten)
- Tools
  - Chamblon: Terminology Extractor 3.0
    - Einfaches Setup; nur Anzeigen, Sortieren, Exportieren; keine Validierung
  - Trados: ExtraTerm
    - Setup-Assistenten; gut parametrierbar; Integration mit MultiTerm iX; auch bilingual; …
  - Xerox: Terminology Suite/TermFinder
    - komplexes Inteface; wenig parametrierbar; liefert Grundform; Validierung in TermOrganizer; auch bilingual; …
- Quelle: K. Fleischmann (2002) http://www.tekom.de/pdf/h02_fp31.pdf

Noise und seine Beseitigung

Manuelle Validierung heisst

- Entsorgen von Termkandidaten mit falscher Klassifikation
- Korrigieren von Termkandidaten mit mangelhafter Segmentierung
  - allenfalls auch Korrigieren/Ergänzen von terminologischer Information wie Fachgebiet etc.



Wieviele Termkandidaten gefunden? Wieviel korrekte Termini stecken darin?

Aufwand

- total: Trados 1,5h, Chamblon 3,0h, Xerox 4,7h
- pro korrektem Terminus: Trados 15s, Chamblon 20s, Xerox 23s – Mensch hat innert 1h 151 Termini gefunden, d.h. ca. 24 t/s

Silence und der Umgang damit

Manuelle Validierung der Vollständigkeit

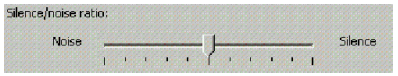
- würde TE-Tools-Verwendung absurd machen

Stichproben sind möglich

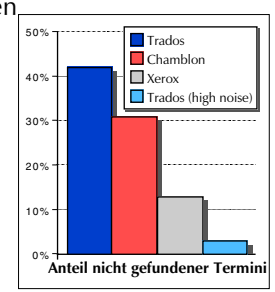
- Studie: Menschlicher Annotator extrahiert während 1h Termini (ca. 20 Seiten Text)

Pragmatische Kompromisse

- Benutzer sollte den Grad an Noise bzw. Silence parametrisieren können



Slider von Trados ExtraTerm



Wieviele Termini hat das TE-Tool übersehen? (bezogen auf Stichprobe)

Evaluationen evaluieren

Wie objektiv sind solche Evaluationen?

- Strenger oder weiter Termbegriff beeinflusst Resultate
  - Eigennamen, Firmennamen, Produktbezeichnungen – »Lufthansa Frequent Flyer Program«, »MacOS X«
  - Abkürzungen, Zahlen, Akronyme, Formeln
  - Anwendungsspezifische textuelle Symbole: Software-Menu-Beschriftungen, Dateinamen, Systembefehle – »fgrep«
- Wie stark werden eingebettete Termini herausgelöst?
  - Sollen in »nuclear power plant« der Terminus »power plant« mitgezählt werden?

Welche Faktoren nebst Silence und Noise machen einzelne TE-Tools mehr oder weniger nützlich?

- Validierungsprozess, Sprachunterstützung, Datenaustausch, Benutzfreundlichkeit, Parametrisierbarkeit, Mehrsprachigkeit, Arbeitsablauf

Linguistische Methoden der TE

Identifiziere spezifisch fachsprachliche Wortbestandteile!

- Affixe, d.h. Präfixe oder Suffixe
  - Medizin: -itis, -aemia, hypo-, peri-

periimunoglobulinaemia

- Stämme
  - Elektrotechnik: -impuls-

Hardwareimpuls
24-Volt-Impulsgeber

Probleme

- direkt nur für einteilige Termini verwendbar
- eng auf Anwendungsbereich abgestimmt (schlecht für kommerzielle Allzweck-TE)

Nebennutzen

- Affixe und Stämme sind oft über Sprachgrenzen hinweg ähnlich und helfen bei der bilingualen Termkandidatenpaarbildung

Hypo-Hyperparathyreoidismus
hypo-hyperparathyroïdisme
hypo-hyperparathyroidis

Linguistische Methoden der TE

Identifiziere Termkandidaten anhand der Wortarten!

- sprachspezifische Wortgruppenmuster für Nominalphrasen, d.h. Nomen mit seinen Attributen

Adjektiv + Nomen — non-financial enterprise
Nomen + Nomen — interbank market
Nomen + »of«-Präposition + Adjektiv + Nomen — settlement of cross-border payments

Problem

- Produziert sehr viel Noise! Insbesondere, wenn Muster "Nomen" für einteilige Termkandidaten zugelassen ist!

Gängiger Ausweg

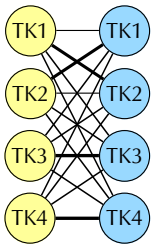
- Verwendung von Stoppwortlisten (manuell erstellt oder Sammlung hochfrequenter allgemeinsprachlicher Vokabeln)

Exkurs Xerox TermFinder

Bilinguale Termextraktion à la Xerox Termfinder

I. Extrahiere monolingual in beiden satzalignierten Paralleltexten alle Nominalphrasen über Wortgruppenmuster als Termkandidaten!

II. Bilde bilinguale Termkandidaten, indem alle monolingualen Termkandidaten aus jedem Parallelsatz miteinander gepaart werden!



Kombinatorische Paarung von Termkandidaten

Problem

- Unzählige falsche Kombinationen!

Ausweg

- Automatisches Ausfiltern schlechter Paare mit Heuristiken (Daumenregeln) zu Übereinstimmung vonTermlänge, interner Struktur etc. ☛

Exkurs Chamblon TerminologyExtractor

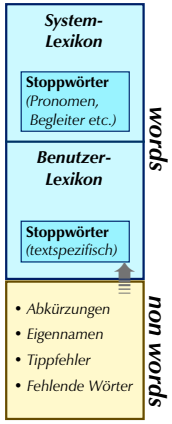
Monolinguale TE

I. Partitioniere die Token der Texte in unbekannte Wortformen (***non words***) und dem Lexikon bekannte Grundformen (***words***)!

- Lexikon besteht aus 4 Komponenten
  - Flexionsauflösendes Systemlexikon (E/F)
  - Eingebaute Stoppwortliste (Funktionswörter)
  - Vom Benutzer erweiterbares Lexikon
  - Vom Benutzer erweiterbare Stoppwortliste

II. Extrahiere die häufigsten *N*-Gramme (Kollokationen) aus ***words*** und ***non words***!

- Reine Stoppwortkollokationen werden ignoriert!
- Eingebettete Kollokationen werden ignoriert, ausser sie sind häufiger als die umfassendere Kollokation!



Partitionierung der Texttoken

Quantitative Methoden der TE

Identifiziere Termkandidaten wegen ihrer abweichenden Vorkommenshäufigkeit!

- **Idee**: Fachwörter kommen in Fachtexten (SL) häufiger vor als in allgemeinen Texten (GL).

- **Relative Häufigkeit** eines Worts in einem Text

$$f_T(w) = \frac{\text{Vorkommen des Worts } w \text{ in Text } T}{\text{Vorkommen aller Wörter in Text } T}$$

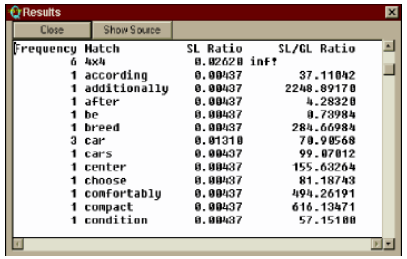
- **Weirdness** eines Worts: Falls hoher Wert, dann Termkandidat!

$$weird(w, SL, GL) = \frac{f_{SL}(w)}{f_{GL}(w)}$$

Quantitative Methoden der TE

Bedingungen

- für TE erst brauchbar bei Vorkommen > 4
- geeignet für einteilige Termini
  - Ausser: Wortgruppen werden als Einheit betrachtet und gezählt!
- relativ sprachunabhängig



Weirdness in SystemQuirk

Dokumentenbezogene Masse

- **Idee**: Fachwörter treten nur in bestimmten Dokumenten auf – dort aber gehäuft!
  - Termini als gute Dokument-deskriptoren (tf/idf aus ***I**nformation **R**etrieval*)

Wie **oft** kommt Term *t* in Dokument *d* vor?
Termhäufigkeit

Wieviele Dokumente enthalten Term *t* (nicht)?
Dokumenthäufigkeit

Quantitative Methoden in der TE

Identifiziere mehrteilige Termkandidaten in Wortgruppen

- **Idee**: Bestandteile von Termini treten auffällig häufig miteinander auf!
  - Assoziationsstärke

Adjektiv + Nomen
non-financial enterprise
W1 W2

- **Ansatz**: Zähle das gegenseitige Vorkommen der Bestandteile in allen Wortgruppen:
  - *a* = W1 und W2 kommen beide vor
  - *b* = W1 kommt ohne W2 vor
  - *c* = W2 kommt ohne W1 vor
  - *d* = weder W1 noch W2 kommt vor

	+W1	-W1
+W2	*a*	*c*
–W2	*b*	*d*

- Einfachstes **Mass** (***S**imple **M**atching **C**oefficient*)
  - Viele weitere Masse; SMC ist aber gut!

$$SMC = \frac{a+d}{a+b+c+d}$$

Quantitative Methoden der TE

Bilinguales Zuordnen von Kandidaten (*term alignment*)

I. Kompilation von bidirektionalen, probabilistischem Lexika (*word alignment*) aus Paralleltexten

II. Alignierung eines Terms *T*
  1. Nimm alle Übersetzungen der Bestandteile von *T*.
  2. Aligniere mit demjenigen Zielterm *Z*, dessen Bestandteile die höchste Summe der Wahrscheinlichkeiten der Übersetzungen von *T* aufweisen.

```
sprememba
------------------
amendments      0.54
changes         0.21
amendment       0.14
Act             0.03
Harmonized      0.02
devices         0.02
medical         0.02
responsibility  0.01
```

Probabilistisches Lexikon: Slowenisch → Englisch (erzeugt mit TWENTE)

Problem

- Seltene Wörter haben schlechte probabilistische Lexikoneinträge
- Zuordnen von einteiligen (Komposita) zu mehrteiligen Termini

Nähere Zukunft – Anwendungen

Bilinguale Termextraktion im Kontext von CAT-Systemen zur technischen Redaktion

- noch engere Verknüpfung mit den Methoden der computergestützten Übersetzung (*translation memories)*
  - Extraktion von Termkandidaten vor der Übersetzung
  - Intelligenter Look-up während der Übersetzung
  - Konsistenzprüfungen nach der Übersetzung
    - *ATA Translation Support Tools Forum* 2002/2003 ☛

Bilinguale Termextraktion für Informationssuche

- CLIR (*cross language information retrieval*)

Monolinguale Terminologieextraktion

- bleibt schwierig: Professionalisierung, Outsourcing

Nähere Zukunft – Methoden

Verbesserte hybride Ansätze

- Das Beste der linguistischen und statistischen Methoden kombinieren
  - Integration von robusten linguistischen Technologien (partielle syntaktische Analyse)
- Hauptziel: Verminderung von Noise, d.h. höhere Präzision
- Erkennen von Termvarianten

Nutzen bestehender terminologischer Bestände

- Intelligente Integration von elektronisch verfügbaren Ressourcen
  - Terminologien, *translation memories*, Thesauri, Ontologien
- Intelligente Interaktion mit Benutzenden

Einbezug semantischer Information