Ausgewählte Techniken der Maschinellen Übersetzung

Susanne J. Jekat

ZHW

E-mail: jes@zhwin.ch, Subject: MTZH

Offene Fragen

- Unterschied zwischen Normalisierung und Lemmatisierung?
- Enthält GermaNet schweizerdeutsche Einträge?
- http://www.ifi.unizh.ch/cl/hess/classes/seminare/semrep/docs/germanet.pdf
- Link zu EuroWordNet Spanisch http://nipadio.lsi.upc.es/wei1.html
- und Link zu Multilingual Isle Lexikcal Entry

http://www.ilc.cnr.it/EAGLES96/isle/complex/frankfurt-hahn_march2004.htm

Übersicht

Themenbereiche:

1. **Erstellung und Standardisierung von Ressourcen für die Maschinelle Übersetzung**
2. Grammatikformalismen für die Maschinelle Übersetzung
3. Maschinelles Dolmetschen
4. Computergestützte Übersetzung
5. Evaluation von Systemen zur Maschinellen Übersetzung

Typen von Ressourcen: Mehrsprachige Sprachdaten

- **Lexikondatenbanken**

- **Terminologiedatenbanken**
- **Textdatenbanken:**
  - **Paralleltexte**
  - **Übersetzte Texte**
  - **Alignments**
- ***?Getaggte Sprachdaten?***
- **Wissensbasen**
- **Standardisierungsbestrebungen, laufende Projekte**
- **praktische Übung**

Thema 1: Erstellung und Standardisierung von Ressourcen für die Maschinelle Übersetzung

Wissensgebiete zu Thema 1:

a) Datenbankkonzepte (Grundlagen)
b) Kennzeichungssprachen
c) Kontrastive Linguistik
d) Lexikographie und Lexikologie
e) MT (Grundlagen)
f) Suchmechanismen und Suchmaschinen
g) Terminologie
h) Termextraktion
i) Texttechnologie
j) Translationswissenschaft? (eher Probleme der Translation)

Ressourcen

Die derzeitigen Ressourcen bestehen aus:

- ein- und mehrsprachigen Lexika mit lexikalischer Semantik
- ein- und mehrsprachigen Terminologiedatenbanken

ein- und mehrsprachigen Textkorpora, annotiert und ggfs. aligniert

Terminologie

Was ist das?

„Gesamtheit der innerhalb eines wissenschaftlichen Systems definierten Fachausdrücke die sich von umgangsprachlicher Verwendung durch exakte Definition innerhalb eines bestimmten Systems unterscheiden.“ Bussmann 2002

s.a. Fachsprache: „Sprachliche Varietät mit der Funktion einer präzisen, effektiven Kommunikation über meist berufsspezifische Sachbereiche und Tätigkeitsfelder.“ Bussmann 2002

Wozu Terminologie in der MT?

Beispiel: „Es sind nur vierrädrige Wagen (Ausnahme Tandem/Random) mit Holz/Stahl o.ä. Speichenrädern (keine Drahtspeichenräder) mit glatter Lauffläche, geschützt durch flaches Eisenband und/oder Gummiauflage, zulässig. Lenkverzögerung und Abweiser sind zulässig, jedoch keine Geschwindigkeits- und Streckenmesser.“*

*Aus LPO 2002:67

Wozu Terminologie in der MT?

Beispiel:

„Es sind nur xxxxx (Ausnahme xxxxx) mit Holz/Stahl o.ä. xxxxx (keine xxxxx) mit glatter xxxxx, geschützt durch flaches xxxxx und/oder xxxxx, zulässig. xxxxx und xxxxx sind zulässig, jedoch keine xxxxx- und xxxxx.“

Mehrsprachige Terminologiedatenbanken

Beispiel:

Eurodicautom (http://europa.eu.int/eurodicautom/Controller)

Terminologiedatenbank der EU

Mehrsprachige Terminologiedatenbanken

Schwerpunkte:

- Extraktion von Fachwortschatz aus Texten, nach linguistischen Kriterien mit statistischen Verfahren zur Bestimmung der Relevanz
- Akquisition von Übersetzungsvorschlägen aus Paralleltexten

Mehrsprachige Terminologiedatenbanken

- Akquisition von definitionsrelevantem Material und Belegen für ontologische Relationen zwischen Begriffen (s.a. Semantic Web)
- Repräsentation multilingualer terminologischer Daten, Aufbau von Thesauri und Wissensbanken
- Terminologische Standardisierung innerhalb von Unternehmen und Definition von kontrollierten Vokabularen (Beispiel *Referent*)

Lexika und Terminologiedatenbanken

trotz traditioneller Unterschiede zwischen Lexikographie und Terminologie sind viele computerlexikographische Verfahren auch für die Terminologie einsetzbar

Textkorpora

Die Annotation umfasst drei weite Kategorien:

a) Identifikation und Markierungen der Textstruktur: Titel, Überschriften, Texteile

wenn schriftsprachliche Texte vorliegen, kann dies fast vollständig automatisch durchgeführt werden

Kategorien der Annotation

b) Identifikation spezifischer Segmente:

Zitate, Sätze, Datum, Namen

c) Hinzufügen von analytischer Information:

Part-of-Speech-Tags, Alignment-Information, prosodische Markierungen, Syntax etc. = teuerste Ressource

b + c mehr manueller Einsatz

Annotation von Texten: Beispiele

a) Markierung der Textstruktur, z. B. LaTeX

\chapter{CAT}\label{chap:CAT}

CAT ist ein interessanter Bereich, den Übersetzer kritisch nutzen sollten.

\section{Nutzen}\label{sec:nutz}

Sie sollten aber wissen, wie wertvoll ihre Beiträge für die Zusammenstellung der Korpora sind.

Annotation von Texten: Beispiele

b) Identifikation von Segmenten, Beispiel VERBMOBIL

<*tGER> ist das weit vom Hauptbahnhof in ~Hannover entfernt ?

- < > = gehört nicht zum Text
- *tGer Turn German
- ~ Tag für Name
- Leerzeichen + Satzzeichen markiert Ende der Äußerung

Part-of-Speech Tag

c) Hinzufügen analytischer Information, Beispiel Part-of-Speech-Tag Wortartenannotation an die sog. Token (Einzelteile des Textes)

- Erst “ Wort “ , im Folgenden “ Token ” .
- Adv $(NN $($, PRP ADJ $(NN $($.

Offenes Problem: korrekte Tokenisierung, z.B. aus wievielen Token besteht DM 240,-?

Alignment-Information

Basis: parallele Texte, d.h. Quelltext und Übersetzung

Alignierung zeigt, welche Teile des Quelltextes durch welche Teile des Zieltextes übersetzt sind

Alignierung auf verschiedenen Ebenen:

Kapitel, Abschnitte, Absätze, Sätze, Phrasen, Wörter (vgl. Folien 11 – 16 aus MT.ppt)

Annotation von Texten

Bedingungen

1. Die den Texten hinzugefügte Information muss ihn verarbeitbar und wiederverwertbar machen.

Beispiel: In einem mehrsprachigen Korpus gesprochener Sprache erfolgt die Markierung aller “hms” als *hm*, (kann aber frz. *hum* oder engl. *uhm* sein).

Annotation von Texten

Neue Verwendung des Korpus, z.B. als Trainingskorpus für einen Spracherkenner, der auch die “hms” korrekt erkennen soll (mögliche Gründe?)

➔ Notwendigkeit der „Wiederaufbereitung“ des Korpus` oder keine Möglichkeit zur Weiterverwendung

Annotation von Texten, Bedingungen

2. der Text darf durch die Annotation nicht verfremdet werden oder die Annotation muss herausgefiltert werden können

Beispiel: also +/eine Standtru=/+ eine Stadt<äh>rundgang oder -rundfahrt würde mich auch interessieren .

Annotation von Texten, Bedingungen

3. wenn das Annotationsschema zu lose ist, ist der Wert der Anwendung fraglich

Beispiel: RST (Rhetorical Structure Theory) zur Beschreibung der Beziehung zwischen den Segmenten eines Textes
(http://www.sil.org/~mannb/rst//rintro99.htm)

RESTATEMENT S (Nebensegment) reformuliert N (Hauptsegment)

ELABORATION S präsentiert zusätzliche Detailinformation zu N

Annotation von Texten, Bedingungen

? 1. A well-groomed car reflects its owner

2. The car you drive says a lot about you.

Restatement oder Elaboration?

Lösung: Warum?

- nach Mann und Thompson 1987 ein Restatement, s.

http://www.sil.org/~mannb/rst/reldefs.htm#Elaborationlinkhttp://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsARestatementRelation.htm

- Gibt es eine Begründung?

HTML

- HTML: Hypertext Markup Language
- Das Format, in dem die meisten Seiten für das WWW geschrieben sind und übertragen werden. Der Text wird mit Auszeichnungselementen (Tags) angereichert.

HTML

- **<HTML>** (Begrenzung für äußerste Schicht der Dokumentstruktur)
- **<HEAD>** (Dokumentenkopf)
- **<TITLE>** Donald T. Duck, Translator**</TITLE>** (Titel des Dokuments in der Titelleiste, wichtig für das Auffinden der Seite)
- **</HEAD>**
- **<BODY>** (Begrenzung für den Körper des HTML-Dokuments)
- **<H1>** Donald T. Duck **</H1>** (Ebene 1)
- **<P>** Ich bin eine sehr nette Ente (Absatzmarkierung)
- **</BODY>**
- **</HTML>**

Abbildung des HTML-Codes

- **Netscape - [Donald T. Duck, Translator]**

- **Donald T. Duck**
- Ich bin eine sehr nette Ente

XML- Extensible Markup Language

- die meisten Dokumente liegen in HTML vor, aber bereits 1994 war absehbar, dass die Grenzen von HTML erreicht waren
- denn: Namen, Anzahl und Kombinationsmöglichkeiten der Auszeichnungselemente, z.B. <table> für eine Tabelle sind vordefiniert und nicht vom Benutzer veränderbar

==> (auch) Schwierigkeiten beim Aufbau effizienter Suchmaschinen

XML- Extensible Markup Language

- aus diesem Grund hat das World Wide Web Consortium (W3C, http://www.w3.org) XML entwickelt
- gestattet explizite Strukturierung beliebiger Informationen, z.B. <strassenadresse>

XML und HTML

- Vielzahl von Auszeichnungssprachen für identische oder ähnliche Einsatzgebiete
- z.B. statt <strassenadresse> <stradr> oder <adr> ==> Überschuss an Struktur

XML und HTML

- Vorteil: Anpassbar und bei geeigneter Entwicklung und Pflege als Standard unterstützend bei Auffinden und Vergleich von Informationen
- z.B. Textkorpora , Text Encoding Initiative (http://www.tei-c.org) hat XML-Dokumenttyp-Definitionen (DTD) zur Auszeichnung von Texten entwickelt. Die DTD enthält Namen von Elementen und Attributen sowie deren Kombinationsmöglichkeiten (Beispiel: das Element s muss das Element np gefolgt von vp enthalten).

XML und HTML

- beide basieren auf SGML, Standard Generalized Markup Language, XML ist Weiterentwicklung von SGML
- SGML-Dokument enthält:

1 . Beschreibung des erlaubten Zeichensatzes und der Zeichen, die zur Unterscheidung des reinen Textes von den Markup-Markierungen verwendet werden

XML und HTML

2 eine Erklärung des Dokumenttyps sowie eine Auflistung der erlaubten Markup-Markierungen, die er enthalten darf, diese sind nicht beliebig kombinierbar

3 das Dokument selbst, dass den Text und die Verweise auf die Markup-Markierungen enthält

- HTML ist in bezug auf 1. und 2. festgelegt ==> in XML annotierte Korpora können in statische HTML-Dokumente konvertiert werden

SGML und XML basierte Informationsmodellierung

- bis zu HTML 4.01 Definition der Namen und der Kombinationsmöglichkeiten der Elemente und Attribute (=DTD) auf der Basis von SGML
- ab 2000 Neuformulierung von HTML auf der Grundlage von XML

vgl. z.B. Pemberton, S. (2002) XHTML1.0: The Extensible Hypertext Markup Language (Second Edition). Technische Spezifikation, W3C.

Quellen

- Bussmann, Hadumod (2002) Lexikon der Spradchwissenschaft. Stuttgart: Kröner..
- Farwell, David (1998) *On Representing Language-specific Information in Interlingua*, http://crl.nmsu.edu/Events/ FWOI/FirstWorkshop/papers.html;
- Fensel, Dieter, Hendler, James, Lieberman, Henry & Wahlster, Wolfgang (2003) Spinning the Semantic Web – bringing the World Wide Web to its Full Potential, MIT Press;
- Harold, E.R. (2001) XML Bible 2nd Edition, New York: Hungry Minds.
- Klabunde, Ralf et al. , ed, (2004) Computerlinguistik und Sprachtechnologie. Eine Eine Einführung. Heidelberg:Spektrum;
- Lobin, H. (2000) Informationsmodellierung in XML und SGML, Berlin: Springer.
- Meyer, Felix (1998) Eintragsmodelle für terminologische Datenbanken: ein Beitrag zur übersetzungs-orientierten Terminographie, Tübingen: Gunther Narr;
- Schwarzl Anja (2001): The (Im)Possibilities of Machine Translation, Frankfurt am Main.