Fortgeschrittene Themen der statistische maschinelle Übersetzung

(SMT II / Advanced SMT)

Übersicht und Einführung

Miriam Kaeshammer
Heinrich-Heine-Universität Düsseldorf

Wintersemester 2014/15

Organisatorisches

Kurszeiten: Dienstag, 10:30 - 12:00

Raum: 23.02.U1.22

Webseite:
`http://user.phil-fak.uni-duesseldorf.de/~kaeshammer/smt14/`

Email: `kaeshammer@phil.uni-duesseldorf.de`

Voraussetzung: abgeschlossener Kurs *Statistische maschinelle Übersetzung* (Sommersemester) oder gleichwertiges Vorwissen

Leistungsnachweis

BN:

- Vortrag zu ausgewähltem SMT-Thema
- Chair für einen anderen Vortrag
- regelmäßige aktive Teilnahme am Kurs
- Bearbeitung von mindestens 75% der Übungen

AP:

- siehe BN
- Klausur (vermutlich in der letzten Semesterwoche)
- Bewertung des Vortrags wird auf Punktzahl der Klausur angerechnet

Vorträge (15-25 Minuten)

Vorstellen einer Erweiterung des phrasenbasierten Modells

Paper-Liste in Google-docs: Eintragen (verbindlich!) für einen Vortrag und einen Vorsitz (Chair) bis einschl. *Mittwoch, 22.10.2014*

Ziele:

- Lesen und Verstehen von wissenschaftlichen Veröffentlichungen
- Anwendung des bis dahin erlernten Grundwissens
- Übermitteln des erarbeiteten Wissens an die Kollegen
- Üben von Vorträgen

Vorbesprechung:

- in der Woche vor(!) dem Vortrag, Terminvereinbarung nach Absprache per Email
- Besprechung (der Struktur) des Vortrags, Klärung von Fragen

Chair

- Vorbereitung: Paper lesen
- Einleitung für den Sprecher
- verantwortlich für die Zeiteinhaltung
- Leitung der Frage- und Diskussionsrunde
- Fragen an den Sprecher

Übungen

Übung/Aufgaben ca. alle zwei Wochen, je nach Ankündigung

Theoretischer & praktischer Teil

Für den praktischen Teil: kleine Implementierungen in Python

→ Voraussetzung: Programmiergrundkenntnisse

Teamarbeit erlaubt, aber jeder gibt eine Lösung ab und gibt an, mit wem er/sie zusammengearbeitet hat

Zusätzliche Literatur

Philipp Koehn, *Statistical Machine Translation*, Cambridge University Press, reprint. with corr., 2011

→ einige Exemplare und als elektronische Ressource in der ULB

Eventuell auch hilfreich:

- C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- Einige Online-Tutorials, z.b. von Kevin Knight (`http://www.isi.edu/natural-language/mt/wkbk-rw.pdf`) oder Michael Collins (`http://www.cs.columbia.edu/~mcollins/notes-spring2013.html`)

Maschinelle Übersetzung

engl. *machine translation*

Abkürzung: MT (oder auch dt. MÜ)

Grundproblem:
gegeben ein Satz in einer *Ausgangssprache* (auch *Quellsprache*),
übersetze automatisch (also mit einem Computerprogramm) in die gewünschte *Zielsprache*

Es gibt viele „richtige“ Übersetzungen!

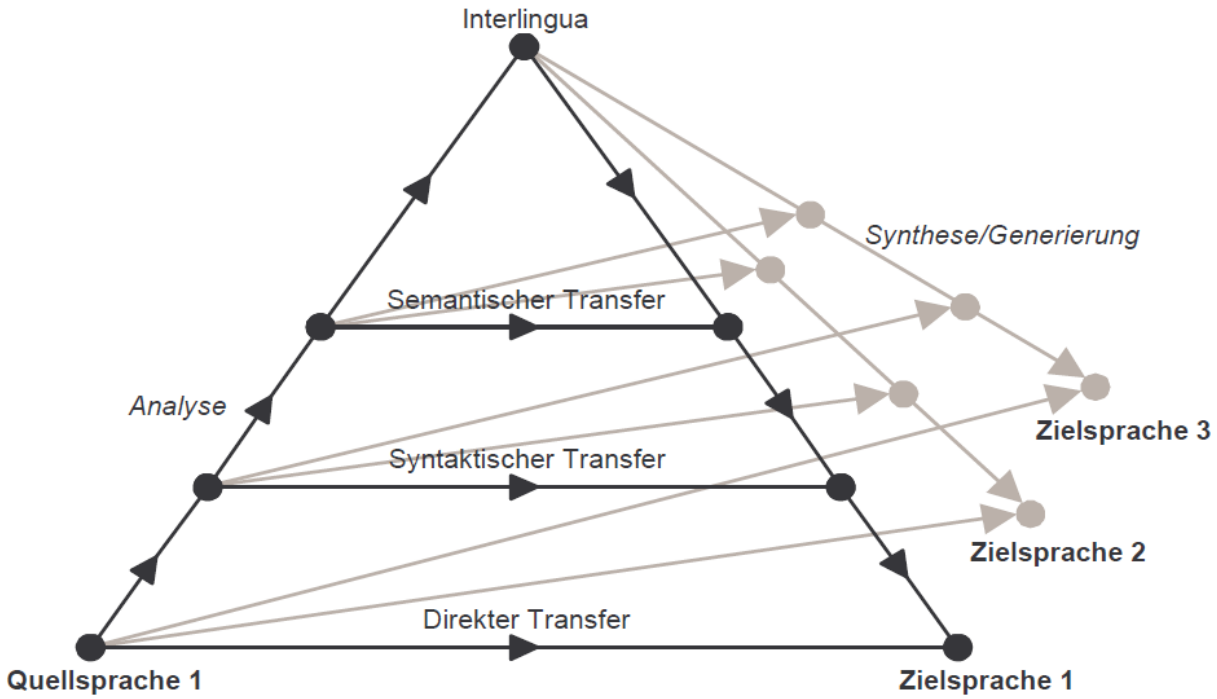
Vorherrschende Paradigmen

1. **Regelbasierte Ansätze**
   - Sprache ist ein begrenztes, regelbasiertes System
   - Automatische Sprachverarbeitung lässt sich mit Regeln definieren
   - Regeln werden anhand von menschlicher Intuition formuliert
2. **Statistische Ansätze**
   - Sprache ist die Gesamtmenge ihrer Instanzen
   - Generalisierungen über Sprache(n) sind möglich auf der Basis von großen Textsammlungen (*Korpora*), die als repräsentative Stichprobe betrachtet werden.
3. **Hybride Ansätze**

Aus Folien von U. Reinke (2005)

Mögliche (regelbasierte) MT-Architekturen

Vauquois-Dreieck/Pyramide



Aus Folien von U. Reinke (2005)

⇒ Pro Sprache ein Analyse- und ein Synthesemodul und pro Sprachpaar ein Transfermodul

Auch: Transfer und Synthese gemeinsam in einem Modul

Regelbasierte Ansätze

Historisch gesehen vor den statistischen Ansätzen

Motivation:

- Gute Übersetzungen setzen linguistisches Wissen voraus, um (a) den Quellsatz zu verstehen und (b) einen wohlgeformten Zielsatz zu generieren.
- (Regelbasierte) Analyse-/Synthesemodule für einige Ebenen (z.B. Morphologie, Syntax) existieren bereits.

Nachteile/Probleme:

- Erfordert viele manuelle Eingaben → hohe Entwicklungskosten
- System wird schnell kompliziert und unübersichtlich → hohe Unterhaltungskosten
- Viel manueller Aufwand für jedes neue Sprachpaar
- Unglaubliche Vielfalt an linguistischen Phänomenen und Übersetzungsäquivalenzen → Abdeckungsproblem

Trotzdem: Erfolgreich in kommerziellen Produkten verwendet (z.B. Systran)

Statistische Ansätze (1)

Idee: Übersetzung durch maschinelles Lernen

1. Entwurf eines *sprachenunabhängigen Modells*, das beschreibt, wie man von einer Sprache in eine andere Sprache übersetzt und jeder möglichen Übersetzung eine Wahrscheinlichkeit/Kosten zuweist.
2. Ein Modell besitzt eine Vielzahl von *Parametern*. Diese werden aus bilingualen Daten (*parallele Korpora*) gelernt. → „Training“
3. Das Modell ist jetzt bekannt – Eingabesätze können nun automatisch übersetzt werden. Es wird ein *Suchalgorithmus* benötigt, der den Zielsatz mit der größten Wahrscheinlichkeit (oder mit den geringsten Kosten) ermittelt. Dieser stellt die Übersetzung dar.

Statistische Ansätze (2)

- Modell muss nur einmal entworfen werden.
- Übersetzungssystem für ein neues Sprachpaar kann schnell trainiert werden; Adaption auf neue Domänen ist ebenso möglich
- Resultate sind natürlich abhängig vom Sprachpaar und der Menge der Trainingsdaten, aber vergleichbar mit denen der regelbasierten Systeme, siehe z.B. Google translate
- Immer mehr und größere parallele Korpora sind verfügbar
- Gegenstand aktiver Forschung

Wie gut ist statistische maschinelle Übersetzung? (1)

French input

Nous savons trés bien que les Traités actuels ne suffisent pas et qu'il sera nécessaire à l'avenir de développer une structure plus efficace et différente pour l'Union, une structure plus constitutionnelle qui indique clairement quelles sont les compétences des États membres et quelles sont les compétences de l'Union.

Statistical machine translation

We know very well that the current treaties are not enough and that in the future it will be necessary to develop a different and more effective structure for the union, a constitutional structure which clearly indicates what are the responsibilities of the member states and what are the competences of the union.

Human translation

We know all too well that the present Treaties are inadequate and that the Union will need a better and different structure in future, a more constitutional structure which clearly distinguishes the powers of the Member States and those of the Union.

Aus Koehn (2010)

System: fr-en, University of Edinburgh (WMT 2005 shared task)

Wie gut ist statistische maschinelle Übersetzung? (2)

Chinese input

伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Statistical machine translation

The London Daily Express pointed out that the death of Princess Diana in 1997 Paris car accident investigation information portable computers, the former city police chief in the offices of stolen.

Human translation

London's Daily Express noted that two laptops with inquiry data on the 1997 Paris car accident that caused the death of Princess Diana were stolen from the office of a former metropolitan police commissioner.

Aus Koehn (2010)

System: cn-en, University of Edinburgh (NIST 2006 campaign)

Ist maschinelle Übersetzung überhaupt nützlich?!

Je besser die Ausgabequalität von maschineller Übersetzungstechnologie, desto nützlicher die Systeme

Aber: auch **maschinelle Übersetzung niedriger Qualität** kann durchaus brauchbar sein

→ abhängig von der Anwendung

Anwendungen von maschineller Übersetzung

Einsatz von (automatischer) Übersetzung

1. ***Aufnahme*, engl. *assimilation***
   Übersetzung eines fremdsprachigen Textes, um den Inhalt zu verstehen
   → Robustheit, Abdeckung

2. ***Verbreitung*, engl. *dissemination***
   Übersetzung eines Textes, um ihn in einer Fremdsprache zu veröffentlichen
   → Qualität

3. **Kommunikation**
   Übersetzung von Emails, Chatroom-Diskussionen, sogar Unterhaltungen (→ Spracherkennung)
   → Geschwindigkeit, Kontextabhängigkeit

Risiken bei der Verbreitung von „übersetzten“ Texten

'I am not in the office at the moment. Please send any work to be translated'

Aus Folien von A. Eisele (2010)

Anwendungen (1)

Vollautomatische, hochwertige maschinelle Übersetzung
engl. fully-automatic high-quality machine translation (FAHQMT)

- Das Übersetzungsproblem ist nur teilweise linguistisch!
  → Welt- und Kontextwissen wird benötigt.
- Bis jetzt nur möglich für *begrenzte Domänen*,
  z.B. Wettervorhersage, Zusammenfassung von Sportereignisse,
  Fluginformationssysteme
- Ausweg: *kontrollierte Sprache*

***Gisting* - Kerninhalte fremdsprachiger Text verstehen**

- Übersetzungsqualität muss nicht perfekt sein
- Anwender: Internetnutzer, Geheimdienste, ...
- Mittlerweile sogar MT zur Verbreitung (siehe
  Microsoft-Hilfeartikel)

Anwendungen (2)

Verbindung mit Sprachtechnologie
Übersetzung von Telefonunterhaltungen, Tonübertragungen usw.

- In Spracherkennung und SMT werden ähnliche Ideen und Modelle verwendet → direkte Kombination möglich
- Sprachübersetzung in Echtzeit ist möglich

Nachbearbeitung, engl. *Post-editing*
Überbegriff: Human-aided machine translation

- Ziel: zur Veröffentlichung geeigneter Text
- Zuerst MT, dann menschliche Nachbearbeitung
- Kann Übersetzungskosten einsparen, wenn der Aufwand für die Nachbearbeitung geringer ist als für die Komplettübersetzung

Anwendungen (3)

Werkzeuge für Übersetzer

interaktive Umgebung für menschliche Übersetzer → höhere Produktivität

- Übersetzungsspeicher, engl. *translation memory*
- MT-System evaluiert sein eigenes Vertrauen in die Übersetzung. Ist dieses zu niedrig → menschliche Übersetzung
- ...

Historischer Abriss (1)

- Die Idee von einer Maschine, die übersetzt, gibt es schon (mindestens) so lange wie elektronische Computer (1940er).
- Zweiter Weltkrieg: Briten benutzen Computer, um die deutsche Enigma-Verschlüsselung zu knacken.
- Warren Weaver, Pionier für die maschinelle Übersetzung:
  When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode". (1947)
  [. . .] statistical semantic studies should be undertaken [. . .] as a first step to solve the translation problem (1949)
- 1950er-1970er: großer Enthusiasmus und Optimismus, viel finanzielle Mittel (Regierungen!), regelbasierte MT-Methoden (Transfer, Interlingua)

Historischer Abriss (2)

- 1966: ALPAC-Report bereitet der allgemeinen Euphorie und dem Investitionswillen ein Ende (hauptsächlich in den USA)
- Nichtsdestotrotz gibt es in den 70er die ersten kommerziellen Übersetzungssysteme: 1976 *Météo*, 1968 Gründung *Systran* (1970 Russisch-Englisch für US Air Force,
  1976 Französisch-Englisch für die Europäische Kommission)
- 1980er/1990er: Forschungstrend Interlingua
- späte 80er/frühe 90er: Einzug von statistischen Methoden in NLP: Spracherkennung (F. Jelinek u.a. @IBM), POS-Tagging, dann auch SMT (@IBM) $\rightarrow$ Übersetzung wird als statistisches Optimierungsproblem formalisiert (Obwohl bahnbrechend, setzt sich der Ansatz nicht sofort durch.)
- 1993-2000, Deutschland: Verbmobil-Projekt (Entwicklung von sowohl Interlingua-System als auch statistischen Methoden)

Historischer Abriss (3)

SMT kommt langsam in Fahrt ...

- 1999: Workshop an der John Hopkins University
  → Open-Source-Reimplementierung der SMT-Methoden von IBM (GIZA)
- Seit 2001: DARPA (Behörde des US-Verteidigungsministeriums) zeigt Interesse an MT und finanziert große MT-Forschungsprojekte & -Evaluationskampagnen.
- Technologische Fortschritte: steigende Rechenleistung, Datenspeicherung, Internetwachstum/Verfügbarkeit von digitalen Texten
- Verschiedene Firmen beschäftigen sich mit SMT: Language Weaver (2002), Google, Microsoft usw.

Historischer Abriss (4)

- 2002: P. Koehn veröffentlich die erste Version von *Europarl*
- 2003: *Statistical Phrase-based Translation* von P. Koehn, F. J. Och und D. Marcu
- Seit 2006: Entwicklung von *Moses*, SMT-Toolkit, als freie Software
- Seit 2006: *EuroMatrix*, EU-Projekt zu MT zwischen allen EU-Sprachen, und weitere MT-bezogene Projekte der EU
- 2007: *Google translate* benutzt eigene SMT-basierte Software (vorher Systran)

Für aktuelle Veröffentlichungen:

- Machine Translation Archive: `http://www.mt-archive.info/`
- ACL Anthology: `http://aclweb.org/anthology-new/`

Vorhandene Resourcen und Systeme

Um ein SMT-System für ein bestimmtes Sprachpaar zu bauen, braucht man:

1. ein allgemeines SMT Toolkit, z.B. Moses
2. einen parallelen Korpus für das Sprachpaar

Freie Software

- **GIZA++**: Implementierung der wortbasierten IBM-Modelle, heute hauptsächlich zur Wortalignierung benutzt
- **Berkeley Aligner**: Wortalignierungssoftware
- **SRILM**, **IRST**: Sprachmodellsoftware
- **Moses**: Implementierung u.a. eines phrasenbasierten Decoders, zusammen mit Software für das Trainieren und Tunen der Modelle und für die Evaluierung
- **Joshua**, **cdec**: weitere Übersetzungssoftware
- **BLEU**, **METEOR**: Software zur Auswertung der Güte eines MT-Systems
- ...

Bilinguale Daten

Frei verfügbar

- **Europarl**: Debatten des europäischen Parlaments, 21 Sprachen, bis zu 2 Millionen Sätze pro Sprache
- **OPUS**: riesige Sammlung von frei verfügbaren, parallelen Korpora, $>$ 90 Sprachen, $>$ 40 Milliarden Tokens

Gegen Bezahlung

- Viele parallele Korpora vorhanden, z.B. für Englisch-Arabisch, Englisch-Chinesisch
- Die meisten sind über das LDC (Lingustic Data Consortium, University of Pennsylvania) zugänglich.

Vorverarbeitung der Daten (1)

Sätze vs. Text

- Hauptgegenstand der Forschung: Übersetzung einzelner Sätze
- Dabei geht oft wichtiger Kontext verloren.

  The window is open. **It** is blue. → La fenêtre est ouverte. **Elle** est bleue.

  He is trying. → Er versucht es. / Er bemüht sich. / Er ist ein anstrengender Mensch.

Vorverarbeitung der Daten (2)

Vorverarbeitung

- **Satzsegmentierung**, engl. sentence segmentation
- **Tokenisierung**, engl. tokenization: Segmentierung eines Textes in Einheiten der Wortebene (Problem: Definition „Wort“)

  Hans-Joachim kauft in New York Fish’n’Chips für $2.50. →

  Hans-Joachim

  kauft

  in

  New York

  Fish’n’Chips

  für

  $

  2.50

  .

Vorverarbeitung der Daten (3)

Optionale Vorverarbeitungsschritte

- Normalisierung der Klein-/Großschreibung
  - Generelle Kleinschreibung
    → weniger Datenknappheit, aber Informationsverlust
  - „Richtige“Klein- und Großschreibung, engl. true-casing
- Erkennung von Zahlen (`zwei`, `2.0`), Datumsangaben, Eigennamen usw.

Nach der Übersetzung: Umkehrung der Schritte

Programme für diese Schritte sind größtensteils frei verfügbar.