Softwareprojektpraktikum
Maschinelle Übersetzung

Matthias Huck, Markus Freitag
{huck,freitag}@i6.informatik.rwth-aachen.de

Vorbesprechung 1. Aufgabe 14. April 2011

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Motivation

Mark Twain, the famous writer, once said

> **“In Paris they just simply opened their eyes and stared when we spoke to them in French! We never did succeed in making those idiots understand their own language.”**

Motivation

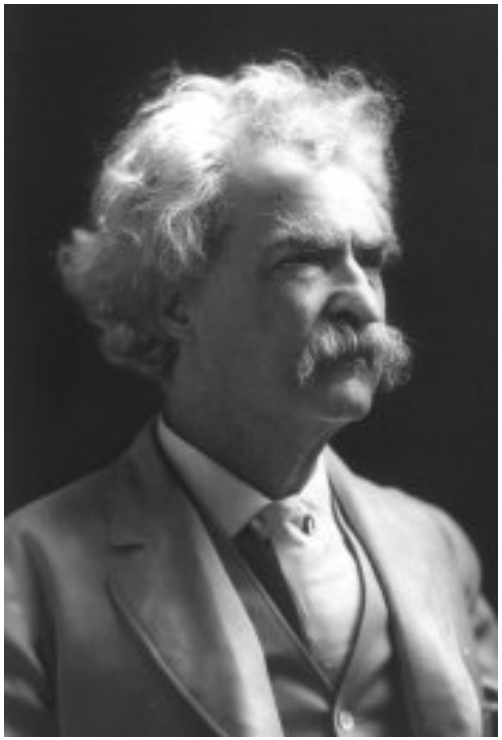
Mark Twain, the famous writer, once said

> **“In Paris they just simply opened their eyes and stared when we spoke to them in French! We never did succeed in making those idiots understand their own language.”**

Mark Twain, der berühmte Verfasser, einmal besagtes

> **“In Paris öffneten sie gerade einfach ihre Augen und starrten an, als wir mit ihnen auf französisch sprachen! Wir nie folgten, mit, jene Idioten zu bilden, verstehen ihr eigenes language.”**
>
> **–babelfish**

Motivation



Mark Twain, the famous writer, once said

> **“In Paris they just simply opened their eyes and stared when we spoke to them in French! We never did succeed in making those idiots understand their own language.”**

Mark Twain, der berühmte Verfasser, einmal besagtes

> **“In Paris öffneten sie gerade einfach ihre Augen und starrten an, als wir mit ihnen auf französisch sprachen! Wir nie folgten, mit, jene Idioten zu bilden, verstehen ihr eigenes language.”**
>
> **–babelfish**

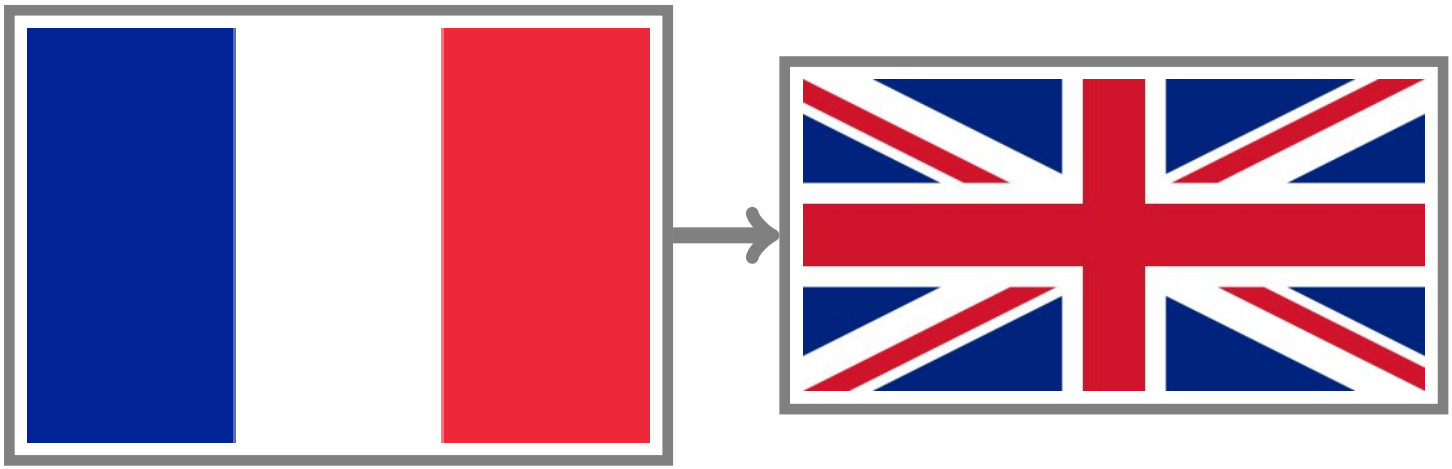
Contents

1 Lehrstuhl i6

Forschung am Lehrstuhl für Informatik 6:

- **Anwendung statistischer Verfahren zur/zum automatischen**
  - **Spracherkennung**
  - **Sprachübersetzung**
  - **Sprachverstehen**
  - **Bilderkennung**
  - **Information Retrieval**
- **angewandte Methoden:**
  - **Mustererkennung**
  - **Signalverarbeitung**
  - **Informationstheorie und statistische Inferenz**
  - **Suchverfahren und effiziente Algorithmen**
  - **Künstliche Intelligenz und Verarbeitung unsicheren Wissens**

2 Einführung



- **Aufgabe in diesem Praktikum:**
  - **Erstellen eines automatischen maschinellen Übersetzers**
  - **Sprachpaar: Französisch–Englisch**
  - **Bewertung und Verbesserung der Übersetzungsresultate**

Das Praktikum

Voraussetzungen:

- **Kenntnisse in Algorithmen und Datenstrukturen und**
- **objektorientierter Programmierung**

Ziele:

- **praktische Erfahrung mit der Programmiersprache C++**
- **praktische Erfahrung in der Programmentwicklung unter Linux**
- **Softwareentwicklung im Team**
- **Implementierung von Datenstrukturen und effizienten Algorithmen**
- **Erwerb von Kenntnissen über Methoden der Sprachverarbeitung**

3 Maschinelle Übersetzung

Ansätze:

- **regelbasiert, knowledge-driven**
  - **bilinguale Sprachexperten erstellen manuell Regeln**
- **statistisch, data-driven**
  - **keine harten Regeln festgelegt**
  - **Computer lernt Sprachzusammenhänge aus Trainingsdaten**
  - **“Siegeszug” der statistischen Übersetzung: seit 1993 (Arbeiten bei IBM)**

Statistischer Ansatz

It must be recognized that the notion of a *probability of a sentence* is an entirely useless one, under any interpretation of this term.

– Noam Chomsky, 1969

- **Gegeben:**
  - **Trainingsdaten, d.h. eine Sammlung von Sätzen der Quellsprache und deren Übersetzung in der Zielsprache**
  - **Beispiel: Reden im Europa-Parlament müssen per Gesetz in alle offiziellen Amtssprachen übersetzt werden**

- **Gesucht:**
  - **Die beste (= wahrscheinlichste) Übersetzung eines unbekannten Satzes**
  - **Bewertungskriterien für die Qualität einer Übersetzung**

Statistischer Ansatz

- **Terminologie:**
  - f **bezeichnet einen Satz in der Quellsprache**
  - e **bezeichnet einen Satz in der Zielsprache**
- **Wahrscheinlichkeitsverteilung** $Pr(e|f)$ **für alle möglichen Übersetzungen** e **eines Quellsatzes** f
- **Finde Zielsatz, der die Wahrscheinlichkeit maximiert:**

$$\hat{e} = \underset{e}{\textbf{argmax}} \{Pr(e|f)\} \tag{1}$$

Bayes

- **Nach Bayes' Entscheidungsregel können wir $Pr(e|f)$ umschreiben als:**

$$Pr(e|f) = \frac{Pr(f|e) \cdot Pr(e)}{Pr(f)} \tag{2}$$

- **Für argmax über alle e ist $Pr(f)$ Konstante**
- **Damit bleibt:**

$$\hat{e} = \underset{e}{\text{argmax}} \{Pr(f|e) \cdot Pr(e)\} \tag{3}$$

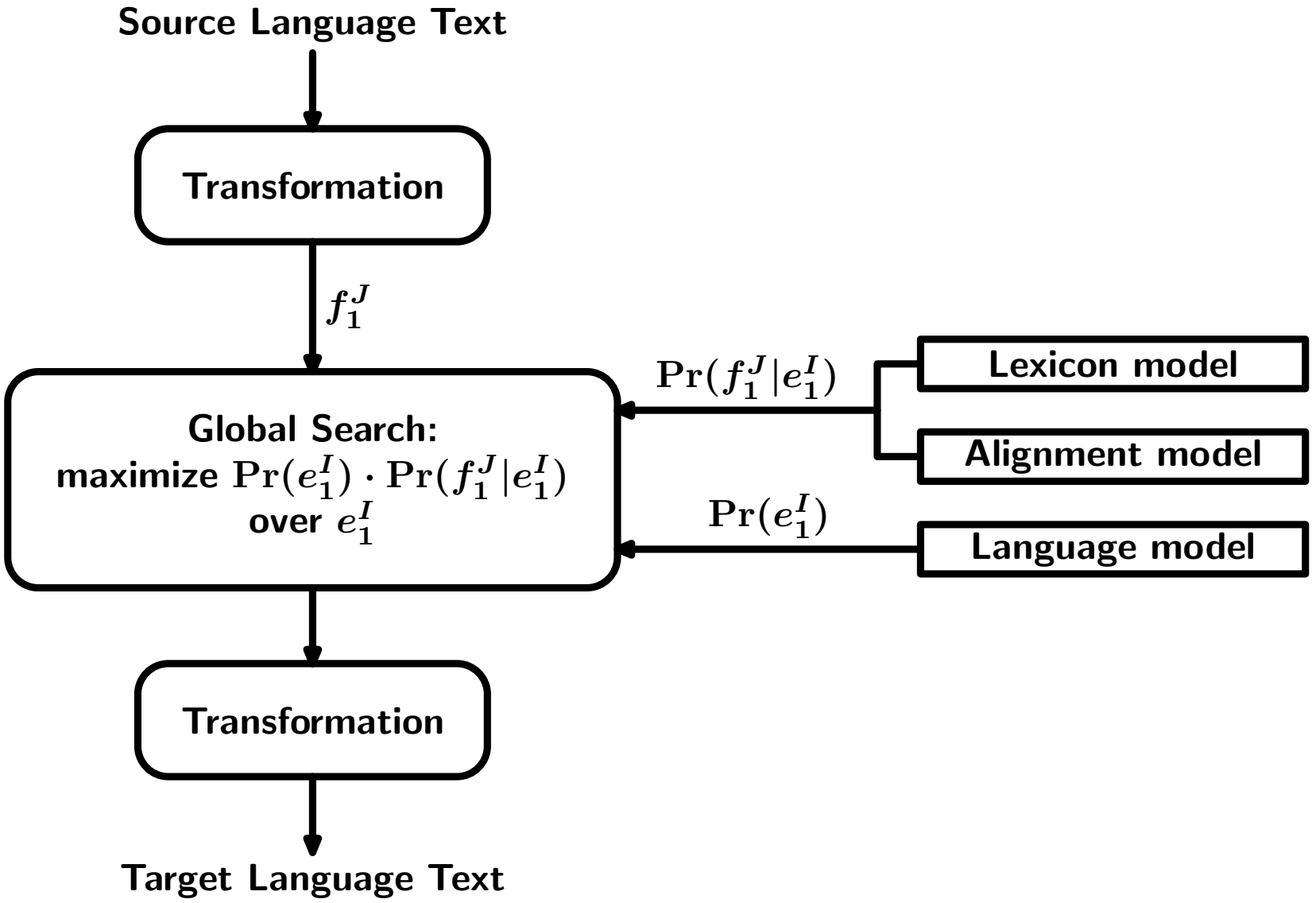
mit

  - **Translation Model $Pr(f|e)$**
  - **Language Model $Pr(e)$**

Teilaufgaben

Teilaufgaben bei der statistischen maschinellen Übersetzung:

- **Training:**
  - **Translation Model**
  - **Language Model**
- **Suche nach der besten Übersetzung**
- **Bewertung der Übersetzungsqualität**
- **Optimierung der Modellgewichtung**
- **Vor-/Nachverarbeitung für typische Fehler**


Source Language Text
Transformation
f_1^J
Global Search:
maximize $\Pr(e_1^I) \cdot \Pr(f_1^J|e_1^I)$
over e_1^I
$\Pr(f_1^J|e_1^I)$
Lexicon model
Alignment model
$\Pr(e_1^I)$
Language model
Transformation
Target Language Text

Translation Model

- **Das Translation Model $Pr(f|e)$ gibt an, wie wahrscheinlich f der Ursprung von e ist**
- **Beispiel für einfache Wörter:**

```
Katze # chat      (75%)
Katze # chatte    (23%)
Katze # mistigri   (2%)
```

- **Vorteil von einzelwortbasierter Übersetzung:**
  - **Flexibel für ungesehene Satzkonstellation**
- **Nachteil von einzelwortbasierter Übersetzung:**
  - **Modellierung größerer zusammengehöriger Einheiten nicht möglich**
  - **Verlust von Kontextinformation**
- **später im Praktikum: längere Phrasen**

Language Model

- **Das Language Model $Pr(e)$ gibt an, wie wahrscheinlich der Satz e in der Zielsprache ist**
- **Üblicherweise als Wahrscheinlichkeit bei $n-1$ Vorgängerwörtern (sog. *n-grams*)**
- **Beispiel für $n = 2$ (Wahrscheinlichkeit gegeben einem Vorgängerwort):**

```
Guten _______
      Tag     (60%)
      Morgen  (25%)
      Mut     (5%)
    ...
```

- **Probleme u. Aufgaben**
  - **Trade-Off zwischen Größe/Genauigkeit**
  - **Handhabung unbekannter Wörter**
  - **Bewertung von Teilübersetzungen**

4 Praktikumsablauf

- sechs Aufgabenblätter
- Gruppen zu jeweils vier Studenten
- Koordination, Schnittstellenverwaltung, Programmierung, Testen

Praktikumsinhalte

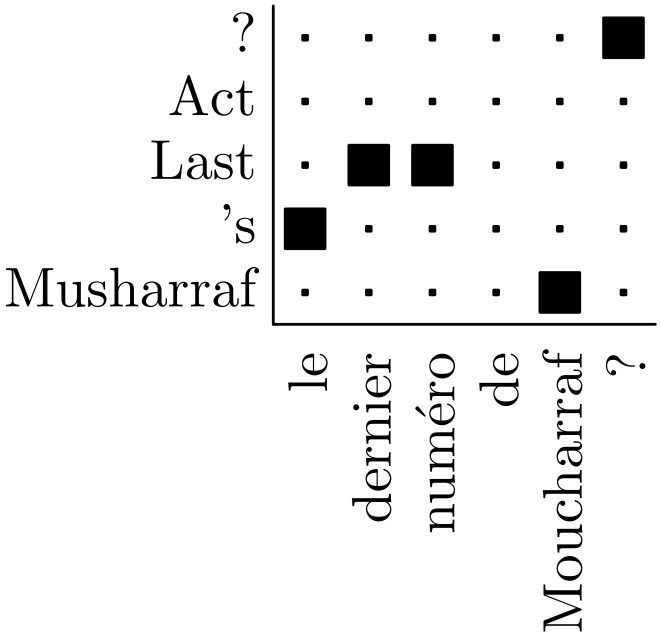
- **Training: Wort-Übersetzungstabelle mit relativen Häufigkeiten** **(Aufgabenblatt 1)**
- **Suche: Suchalgorithmus auf Wortebene** **(Aufgabenblatt 2)**
- **Bewertung: Fehlermaße WER, PER und BLEU** **(Aufgabenblatt 3)**
- **Training II: Präfixbaum und Phrasenübersetzungstabelle** **(Aufgabenblatt 4)**
- **Suche II: Erweiterung des Suchalgorithmus auf Phrasen** **(Aufgabenblatt 5)**
- **Sprachmodellierung: Erzeugen eines Bigramm-Sprachmodells, Verwendung im Rescoring auf n-best-Listen** **(Aufgabenblatt 5)**
- **Modellkombination: Log-lineare Modellierung** **(Aufgabenblatt 6)**
- **Optimierung: Downhill-Simplex Algorithmus, Minimum Error Rate Training** **(Aufgabenblatt 6)**
- **Zusätzlich: Aufgaben zur Software-Architektur**

Aufgabe 1

- **Gegeben: 84189 Satzpaare Französisch – Englisch**
- **Zuordnung zwischen den Wörtern (sog. Alignment), Format:**

```
SENT: 0
S 0 1
S 1 2
S 2 2
S 4 0
S 5 4

SENT: 1
S 1 0
S 2 1
...
```



Alignment

- **Indizes fangen bei Null an**

Aufgabe 1

- **Auslesen der Dateien (gzip-Format, Klasse gzstream.cpp erforderlich)**
- **Erstellen eines Alphabets** `string` $\longrightarrow$ `integer`
- **Berechnung der relativen Häufigkeiten:**

$$p(e|f) = \frac{N(e,f)}{N(f)} \tag{4}$$

$$p(f|e) = \frac{N(e,f)}{N(e)} \tag{5}$$

- **Numerisch stabiler: negative Logarithmen**
- **Erstellen eines Makefiles**

Organisatorisches

Einführung in die Aufgaben und Ausgabe der Aufgabenblätter:

- zweiwöchentlich im Seminarraum des Lehrstuhls i6 (Raum 6124)
- wann? donnerstags, 16:00 Uhr
- Termine: voraussichtlich 14.04. / 28.04. / 12.05. / 26.05. / 09.06. / 30.06.

Reservierung des Rechnerpools zur Bearbeitung der Aufgaben:

- alle Rechner im lila Raum (4U15)
- mittwochs, 16:00 - 20:00 Uhr, und donnerstags, 12:00 - 16:00 Uhr

Kontrolle der Lösungen im Rechnerpool jeweils zwei Wochen nach Ausgabe:

- wann? donnerstags vor der Ausgabe bzw. nach Terminabsprache mit den betreuenden Assistenten
- spätestens 18:00 Uhr des Vortages Abgabe der Lösungen per E-Mail
- Termine: voraussichtlich 28.04. / 12.05. / 26.05. / 09.06. / 30.06. / 14.07.

Sonstiges

- **Zugangsberechtigung im Rechnerpool besorgen**
- **Website:**
  `http://www-i6.informatik.rwth-aachen.de/web/Teaching/LabCourses/SS11/Softwareprojektpraktikum/`
- **E-Mail-Verteiler:**
  `mtsoftprak11@i6.informatik.rwth-aachen.de`
- **Fragen & Probleme:**
  möglichst per E-Mail an
  `{huck,freitag}@i6.informatik.rwth-aachen.de`
  oder persönlich bei uns im Büro vorbeikommen (Räume 6126 und 6125)
- **Bachelorstudiengang:**
  Rücktritt von der Veranstaltung ohne Anrechnung eines Fehlversuchs bis maximal drei Wochen nach Veranstaltungsbeginn möglich
- **Praktikum ist ideale Grundlage für spätere Hiwi-Tätigkeit am Lehrstuhl**

Bitte erwägen Sie sorgfältig, wenn Sie jede Frage haben, links, weil jetzt sein konnte eine leuchtende Zeit, um sie zu bitten

Bitte erwägen Sie sorgfältig, wenn Sie jede Frage haben, links, weil jetzt sein konnte eine leuchtende Zeit, um sie zu bitten

Fragen?