Wie funktioniert eigentlich Google Translate?

Utz Pöhlmann

Universität Hamburg - Proseminar: Künstliche Intelligenz

4 poehlma@informatik.uni-hamburg.de

2. August 2015

• mehr als 7000 Sprachen weltweit

- mehr als 7000 Sprachen weltweit
- Reisen sind "IN"

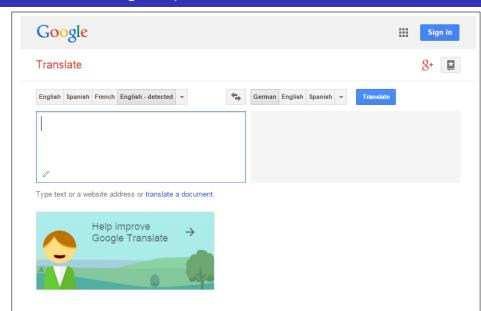
- mehr als 7000 Sprachen weltweit
- Reisen sind "IN"
- Übersetzung überall und für jedermann

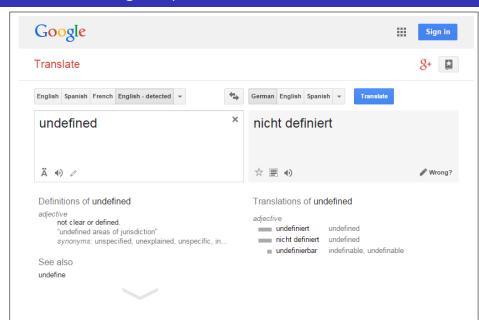
Gliederung

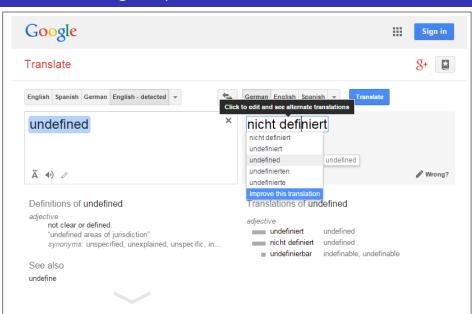
- Einleitung
- Sprachübersetzung von Grund auf
- Schablonentechnik
- Fazit & Ausblick

WLAN?

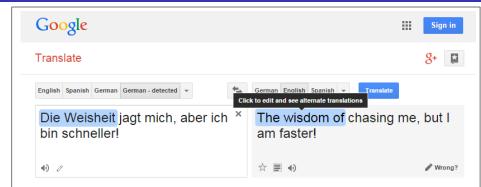
WLAN ist vorhanden! WLAN ist nicht vorhanden!

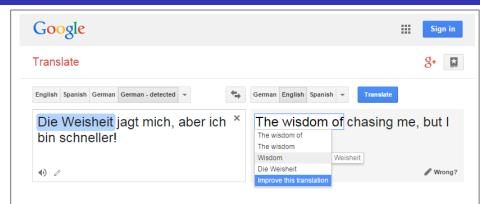




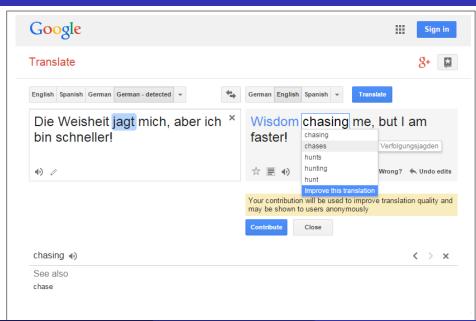




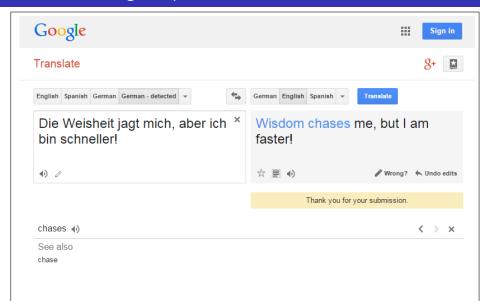












Arten von Übersetzungssystemen

Wort für Wort

Arten von Übersetzungssystemen

Wort für Wort

Schablonentechnik

Arten von Übersetzungssystemen

- Wort f
 ür Wort
 - Übersetzen Marke Lexikon
 - Sinnabschnitte und Wahrscheinlichkeiten
- Schablonentechnik

• Gehe zum ersten Wort

Vorsicht ist die Mutter der Porzellankiste.

- Gehe zum ersten Wort
- Schau ins Lexikon

```
die Vorsicht — attention
die Vorsicht — caution
die Vorsicht — wariness
Vorsicht! — Watch out!
[...]
```

- Gehe zum ersten Wort
- Schau ins Lexikon
- Nimm die erste Wortbedeutung

attention ist die Mutter der Porzellankiste.

- Gehe zum ersten Wort
- Schau ins Lexikon
- Nimm die erste Wortbedeutung
- Gehe zum nächsten Wort

attention ist die Mutter der Porzellankiste.

- Gehe zum ersten Wort
- Schau ins Lexikon
- Nimm die erste Wortbedeutung
- Gehe zum nächsten Wort
- Fahre fort mit Schritt 2

attention actual amount that mother that Porzellankiste.

Befindet sich ein Wort nicht im Lexikon, führe folgende Schritte durch:

Porzellankiste — ?

Befindet sich ein Wort nicht im Lexikon, führe folgende Schritte durch:

 Benutze ein einsprachiges Wörterbuch



Abbildung: Ein beispielhaftes einsprachiges Wörterbuch Quelle: Google

Befindet sich ein Wort nicht im Lexikon, führe folgende Schritte durch:

Porzellankiste = Porzellan + Kiste

- Benutze ein einsprachiges Wörterbuch
- Zerlege das Wort in Teilworte die im Wörterbuch stehen

Befindet sich ein Wort nicht im Lexikon, führe folgende Schritte durch:

- Benutze ein einsprachiges Wörterbuch
- Zerlege das Wort in Teilworte die im Wörterbuch stehen
- Übersetze diese Teilworte

```
das Porzellan — china
[...]
die Kiste — box
[...]
```

Befindet sich ein Wort nicht im Lexikon, führe folgende Schritte durch: attention actual amount that mother that china box.

- Benutze ein einsprachiges Wörterbuch
- Zerlege das Wort in Teilworte die im Wörterbuch stehen
- Übersetze diese Teilworte
- Fahre fort mit dem nächsten Wort

Parallelen

- Englisch und Deutsch beides germanische Sprachen
- also ähnliche Grammatik und ähnlicher Wortschatz

- Parallelen
- Worttypen

Vorsicht – Substantiv

ist – Verb

die – Artikel

Mutter — Substantiv

der – Artikel

Porzellankiste – Substantiv

- Parallelen
- Worttypen
- Voraussagen

Beziehungen:

Vorsicht – keine erkennbar

ist — zu Substantiv

die – zum nächsten Wort

Mutter — keine erkennbar

der – zum nächsten Wort

Porzellankiste – keine erkennbar

- Parallelen
- Worttypen
- Voraussagen
- Filterung

Bisher keine sinnlosen Daten

- Parallelen
- Worttypen
- Voraussagen
- Filterung
- Auftrennen in Wortstamm und Affix

```
Vorsicht – Stamm
```

$$Porzellankiste-Stamm+Stamm$$

- Parallelen
- Worttypen
- Voraussagen
- Filterung
- Auftrennen in Wortstamm und Affix
- Übersetzen unter Beachtung der Wortstellung und der Fälle

Carefulness is the crate of porcelain's mother.

 Textabschnitte in anderen Texten wiederfinden

- Vorsicht ist die Mutter der Porzellankiste.
- Vorsicht ist die Mutter der Porzellankiste.
- Vorsicht ist die Mutter der Porzellankiste.
- [...] Vorsicht ist die Mutter der Porzellankiste [...]

- Textabschnitte in anderen Texten wiederfinden
- Übersetzung aus dem anderen Text nehmen

- Discretion is the better part of valor
- Prudence is the better part of valor.
- Prudence is the better part of valour.
- [...] caution is the mother of wisdom [...]

- Textabschnitte in anderen Texten wiederfinden
- Übersetzung aus dem anderen Text nehmen
- Als Schablone behalten

- 1 = Discretion is the better part of valor
- 2 = Prudence is the better part of valor.
- 3 = Prudence is the better part of valour.
- 4 = caution is the mother of wisdom

- Textabschnitte in anderen Texten wiederfinden
- Übersetzung aus dem anderen Text nehmen
- Als Schablone behalten
- Wenn Schablone ohne gegenseitigen Konflikt den ganzen Text abdecken

Die Schablone steht in Konflikt zu:

Schablone 1 : 2, 3, 4

Schablone 2 : 1, 3, 4

Schablone 3 : 1, 2, 4

Schablone 4 : 1, 2, 3

Schablonen die ohne Konflikt den gesamten Text abdecken:

- 1
- 2
- 3
- 4

- Textabschnitte in anderen Texten wiederfinden
- Übersetzung aus dem anderen Text nehmen
- Als Schablone behalten
- Wenn Schablone ohne gegenseitigen Konflikt den ganzen Text abdecken
 - Diese Schablonen als aktuelle Übersetzung ausgeben
- Sonst kleinere
 Schablonen wählen

Schablone 2 kommt am häufigsten in anderen Schablonen vor.

Deshalb wird sie ausgegeben:

Prudence is the better part of valor.

Aussicht

• Google Translate arbeitet mit diesem Muster

Aussicht

- Google Translate arbeitet mit diesem Muster
- andere Systeme sind schlechter im Übersetzen von ganzen Texten als Google Translate

Aussicht

- Google Translate arbeitet mit diesem Muster
- andere Systeme sind schlechter im Übersetzen von ganzen Texten als Google Translate
 - Babelfish
 - Leo
 - Linguee (Gibt nur Schablonen aus, wagt keinen Übersetzungsversuch)
 - kleine, die es nicht ins Internet geschafft haben
 - und Weitere

• Google Translate bisher beste Lösung

- Google Translate bisher beste Lösung
- Eine gänzlich neue Lösung wurde bisher noch nicht gefunden

- Google Translate bisher beste Lösung
- Eine gänzlich neue Lösung wurde bisher noch nicht gefunden
- Kleinere Bugfixes bestimmt

- Google Translate bisher beste Lösung
- Eine gänzlich neue Lösung wurde bisher noch nicht gefunden
- Kleinere Bugfixes bestimmt
- Wer selber Übersetzen will, nimmt Linguee

Quellenverzeichnis I



Utz Pöhlmann (2015)

Wie funktioniert eigentlich Google Translate?

Entwicklungsstadium Stand Juli 2015.



Front Runner (2015)

Gemeinsamkeiten und Unterschiede der deutschen und englischen Sprache http://front-runner.de/de/gemeinsamkeiten-unterschiede-englisch-deutsch.html Absatz 2, Juli 2015.

Ende

Ende

Noch Fragen?

Parallelen

• Gemeinsamkeiten Ausgangs-, Zielsprache

Parallelen

• Gemeinsamkeiten Ausgangs-, Zielsprache

• Suchen: z.B. Google

Parallelen

- Gemeinsamkeiten Ausgangs-, Zielsprache
- Suchen: z.B. Google
- Besser: vorher einprogrammiert

Substantiv

- Substantiv
 - immer groß geschrieben
 - haben oft Artikel
 - deklinierbar

- Substantiv
 - immer groß geschrieben
 - haben oft Artikel
 - deklinierbar
- Verb

- Substantiv
 - immer groß geschrieben
 - haben oft Artikel
 - deklinierbar
- Verb
 - meist klein geschrieben
 - konjugierbar

- Substantiv
 - immer groß geschrieben
 - haben oft Artikel
 - deklinierbar
- Verb
 - meist klein geschrieben
 - konjugierbar
- Adjektiv

- Substantiv
 - immer groß geschrieben
 - haben oft Artikel
 - deklinierbar
- Verb
 - meist klein geschrieben
 - konjugierbar
- Adjektiv
 - meist klein geschrieben
 - steigerbar
- und weitere



• Beziehung von 2 Wörtern

- Beziehung von 2 Wörtern
 - zusammengehörig (z.B. Artikel + Substantiv)
 - unabhängig

- Beziehung von 2 Wörtern
 - zusammengehörig (z.B. Artikel + Substantiv)
 - unabhängig
- Beziehung von 2 Sätzen

- Beziehung von 2 Wörtern
 - zusammengehörig (z.B. Artikel + Substantiv)
 - unabhängig
- Beziehung von 2 Sätzen
 - Nebensatz
 - implizite Referenzen (z.B. Thema)
 - unabhängig

Filterung

Berechnung aller Daten

Filterung

- Berechnung aller Daten
 - Wortanzahl
 - Abstände von Wörtern
 - etc.

Filterung

- Berechnung aller Daten
 - Wortanzahl
 - Abstände von Wörtern
 - etc.
- Schlechte Ergebnisse verwerfen

Auftrennen in Wortstamm und Affix

• Auftrennung einzelner Wörter

Auftrennen in Wortstamm und Affix

- Auftrennung einzelner Wörter
 - 2 Bäume bauen
 - Wort von hinten und vorne durch gehen
 - gefundene Affixe entfernen

Auftrennen in Wortstamm und Affix

- Auftrennung einzelner Wörter
 - 2 Bäume bauen
 - Wort von hinten und vorne durch gehen
 - gefundene Affixe entfernen
- gefundene Wortstämme auf Existenz prüfen (Lexikon)
- gefundene Affixe auf Existenz pr

 üfen (Datenbank)

• erstes Wort nehmen

- erstes Wort nehmen
- Wortstamm im Lexikon suchen
 - bei mehreren Möglichkeiten Voraussage beachten

- erstes Wort nehmen
- Wortstamm im Lexikon suchen
 - bei mehreren Möglichkeiten Voraussage beachten
- mit Affixen richtige Form finden

- erstes Wort nehmen
- Wortstamm im Lexikon suchen
 - bei mehreren Möglichkeiten Voraussage beachten
- mit Affixen richtige Form finden
- nächstes Wort übersetzen

Krasser Stuff der mir eventuell noch einfällt.