Computerlinguistik

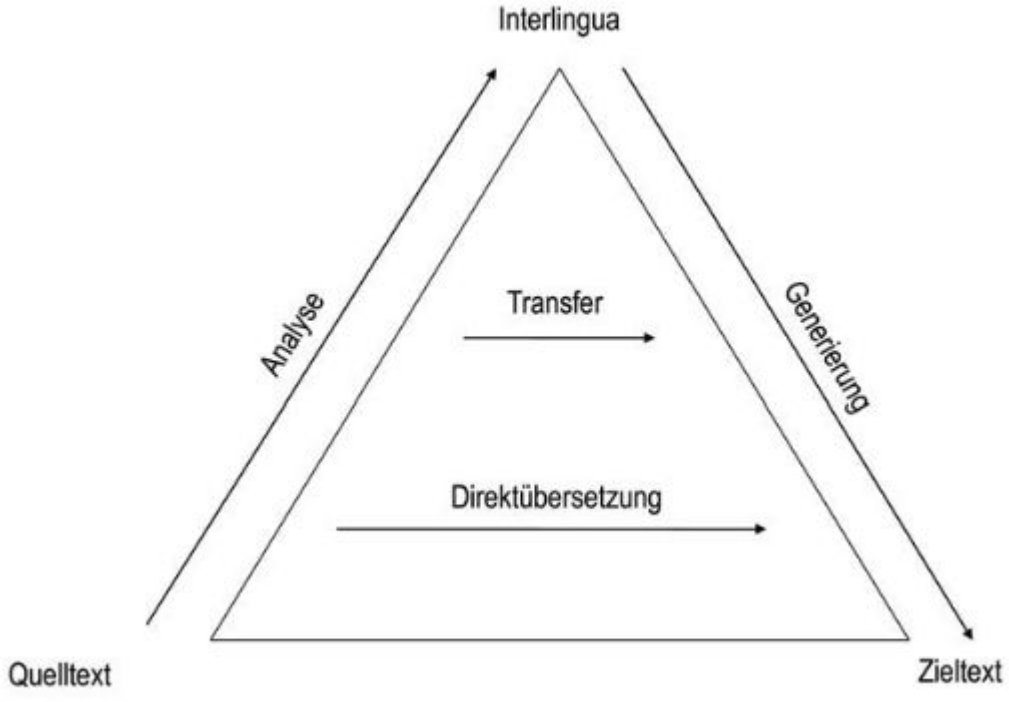
E15: Evaluation

IDH INSTITUT FÜR DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

(Wiederholung) Maschinelle Übersetzung

- Weaver-Memorandum (1949)
- ALPAC-Report (1966)
- EUROTRA (1978 - 1992)
- VERBMOBIL (1993-2000)
- Google Translate (seit 2006)
- DeepL (seit 2017)

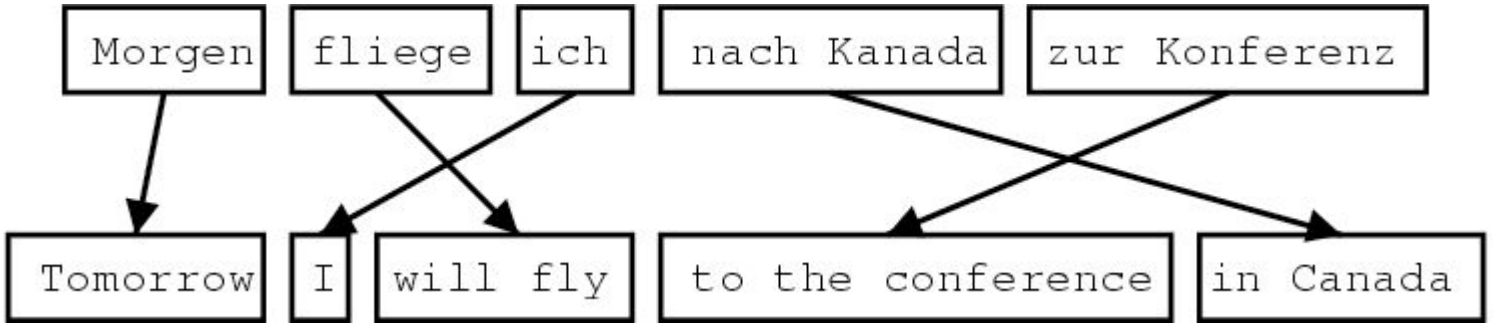
Maschinelle Übersetzung – Regelbasiert



Grafik aus Stein (2009): “Maschinelle Übersetzung - Ein Überblick”(JLCL 24:3)

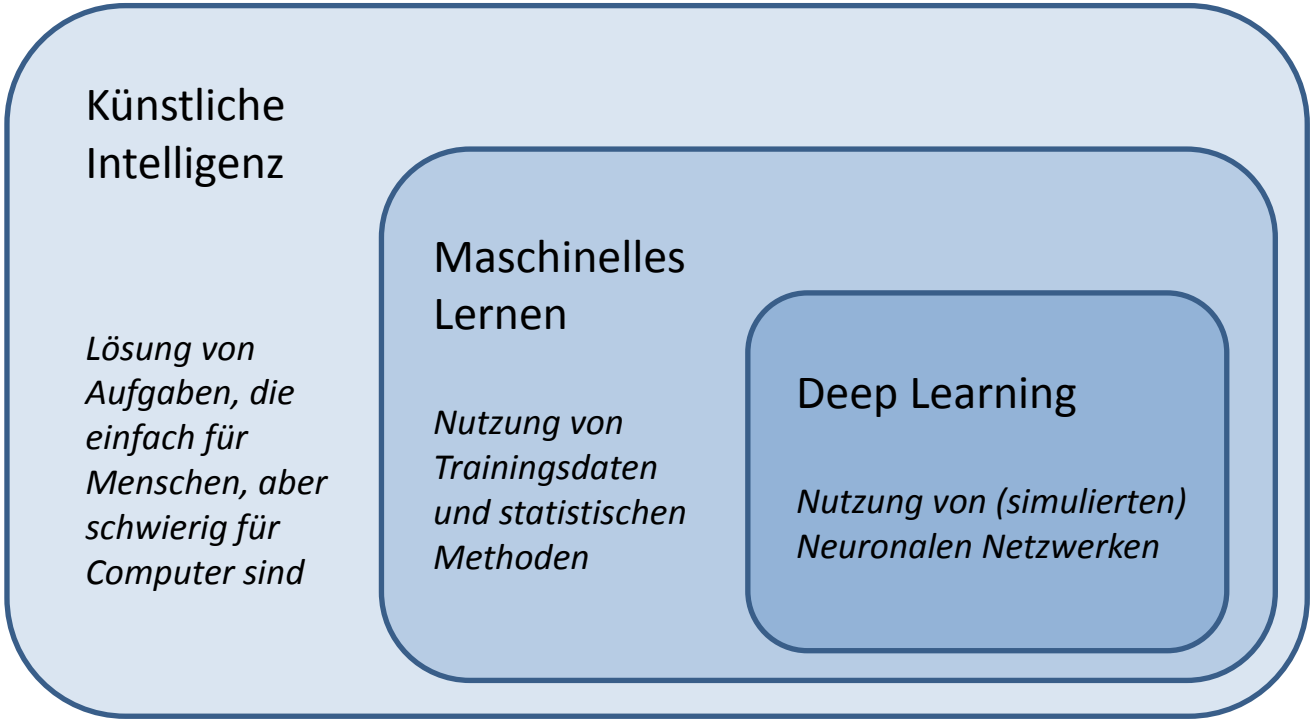
Maschinelle Übersetzung – Statistisch

- Basis: Parallele Korpora mit Alignment



- Modellbildung über Wort- und Phrasenübersetzungen
- Wahl der wahrscheinlichsten Alternative

Maschinelle Übersetzung – Deep Learning



Evaluation: Grundbegriffe

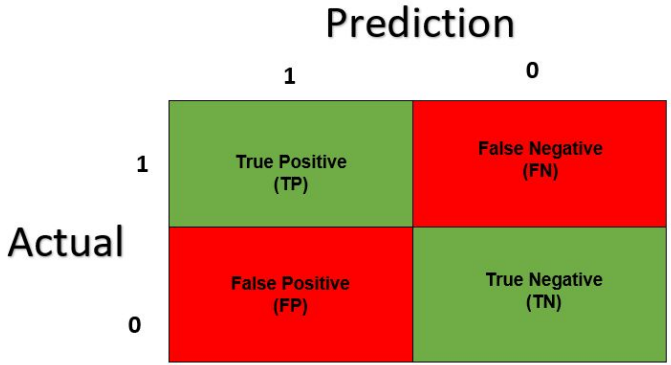
- Zuordnung eines Wertes zu einem Gegenstand (Bewertung)
- Bewertung kann über Messung erfolgen. Diese sollte
  - valide sein → Sie sollte das messen, was sie zu messen vorgibt
  - reliabel sein → Die Messung sollte reproduzierbar sein
- Gründe für die Evaluation:
  - Systemvergleich (meist benutzungsorientiert)
  - Systemverbesserung (meist entwicklungsorientiert)
- Kriterien (ISO/IEC 9126): Funktionalität, Zuverlässigkeit, Usability, Effizienz, Änderbarkeit, Übertragbarkeit

Evaluation verschiedener Bereiche

- Methoden / Theorien:
  - Entsprechen die beobachteten Daten den theoretischen Voraussagen?
- Ressourcen:
  - FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable)
- Anwendungen:
  - Intrinsische Evaluation (von einzelnen Bestandteilen der Anwendung)
  - Extrinsische Evaluation (des Gesamtsystems, Nutzerinteraktion)

Verschiedene Maße für verschiedene Tasks

- Klassifikationsaufgaben:
  - Accuracy $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
  - Precision $\text{Precision} = \frac{tp}{tp + fp}$
  - Recall $\text{Recall} = \frac{tp}{tp + fn}$
  - F-Score $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
- Information Retrieval:
  - Mean reciprocal rank $\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$
- Spracherkennung / Maschinelle Übersetzung:
  - Minimale Editierdistanz
  - Teilbaumübereinstimmung
- ...



		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

```
l e v   e n s h t e i n
o = o + = = = - = = = = =
m e i l e n s   t e i n
```

```
l e   v e n s h t e i n
o = + o = = = - = = = = =
m e i l e n s   t e i n
```

Beispiel: Ein Spamfilter

- Positive: Spam
- Negative: Ham
- Filter bekommt 10000 Mails zur Beurteilung
  - Klassifiziert 5000 als Spam
  - Manuelle Nachkontrolle:
    - 2 der Mails waren gar kein Spam
    - 100 Spam-Mails wurden nicht als solche klassifiziert
- Gesucht: Precision, Recall, F1-Score, Accuracy

Beispiel 2: Achtung, Bayes!

- Rechtschreibprüfprogamm
  - Erkennt 95% der fehlerhaften Wörter
  - Klassifiziert 2% der richtigen Wörter als fehlerhaft
- Wie hoch ist der Anteil der wirklich fehlerhaften Wörter an den beanstandeten Wörtern, wenn nur alle 500 Wörter ein Fehler gemacht wird?

Literatur / Hausaufgabe

- **Zur Nachbereitung**: Lösen Sie folgende Aufgaben (Abgabe über ILIAS):
  - Eine Anwendung soll aus einem Gedichtband alle genannten Tierbezeichnungen finden. Sie liefert insgesamt 55 Terme, von denen 6 keine Tierbezeichnungen sind. Sie lesen jetzt selbst noch einmal die Gedichte und finden insgesamt 57 Tierbezeichnungen. Ermitteln Sie Precision und Recall des Systems. Welche Angabe würden Sie noch benötigen, um die Accuracy des Systems zu ermitteln?
  - Hausaufgabenphobie (HAP) ist eine relativ seltene Erkrankung, sie betrifft nur ungefähr jede|n 250. Studierende|n. Es wurde ein neuer Test (HAPT2) entwickelt, der HAP relativ sicher erkennt – nämlich in 99 Prozent aller Fälle. HAPT2 arbeitet außerdem relativ präzise – von 100 Studierenden, die gar nicht unter HAP leiden, werden nur zwei fälschlicherweise mit HAP diagnostiziert.
    Wie groß ist die Wahrscheinlichkeit, dass ein|e Studierende|r unter HAP leidet, wenn dies vom Test diagnostiziert wurde?