inovex

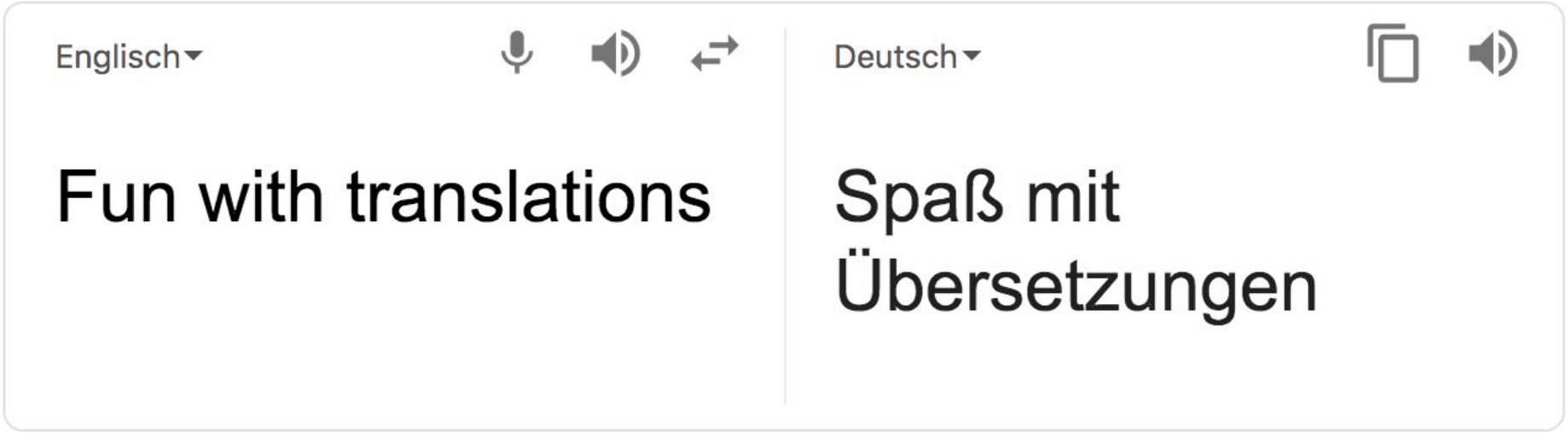
NLP

Maschinelle Übersetzung

Anna Weißhaar & Tilman Berger

Karlsruhe, 29. November 2019

Tilman und Anna präsentieren



inovex

„Es ist die Aufgabe des Übersetzers, jene reine Sprache, die unter dem Zauber einer anderen liegt, in die eigene Sprache freizusetzen, die Sprache, die in einem Werk gefangen ist, in seiner Neuerschaffung jenes Werkes zu befreien.“

(Walter Benjamin)

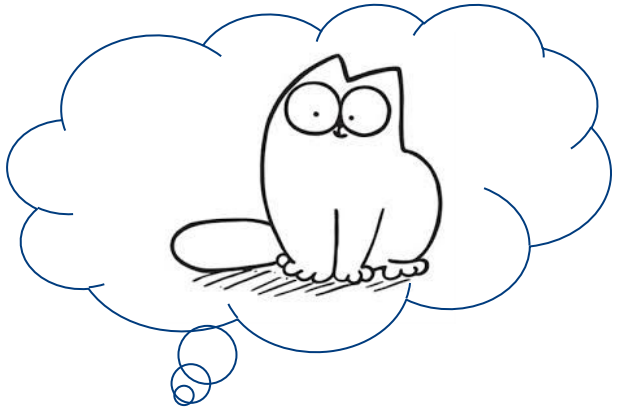
Lost in translation

Zitat von Walter Benjamin:

- Andere Perspektive auf Übersetzung als wir im Lab
- Ähnlicher Grundgedanke:

  "Kern"/(abstrakter) Gedanke eines Textes wird je nach Sprache unterschiedlich ausgedrückt.

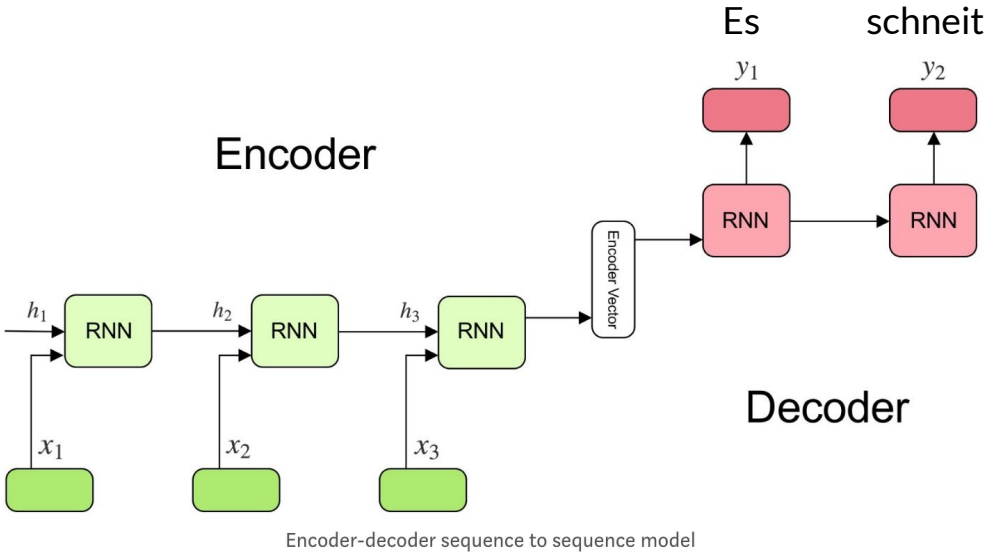
"Katze" =>



=> "cat"

Übersetzung aus ML-Sicht

Many-to-many sequence prediction



It is snowing

https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

inovex

Encoder-Decoder

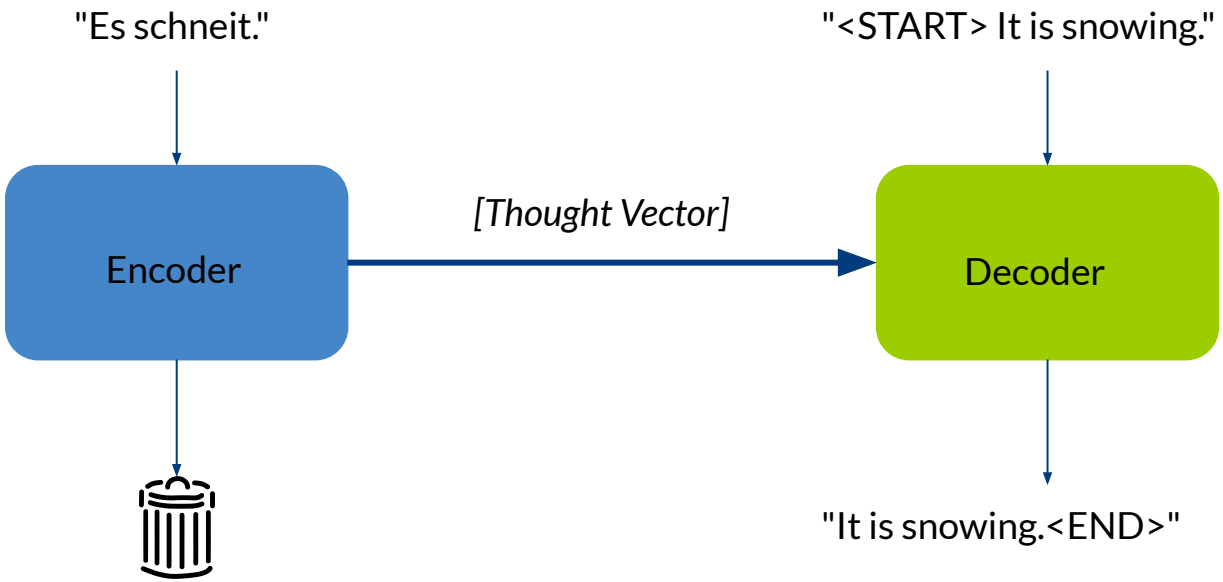
… RNN Encoder-Decoder, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a ***variable-length source*** *sequence to a* ***fixed-length vector****, and the decoder maps the vector representation back to a* ***variable-length target*** *sequence.*

(Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation)

Encoder-Decoder in Stichworten

- Encoder verwandelt Eingabe variabler Länge in *thought vector*.
  - Ausgabe des Encoders wird verworfen.
  - Nur innerer Zustand des Encoders interessant (vgl. Embeddings)
- Decoder wird mit *thought vector* initialisiert und erzeugt daraus Ausgabe ebenfalls variabler Länge.
- "Sequence Embedding": *Thought vector* hat feste Länge.
- Eingabe- und Ausgabelänge können sich unterscheiden.

Training

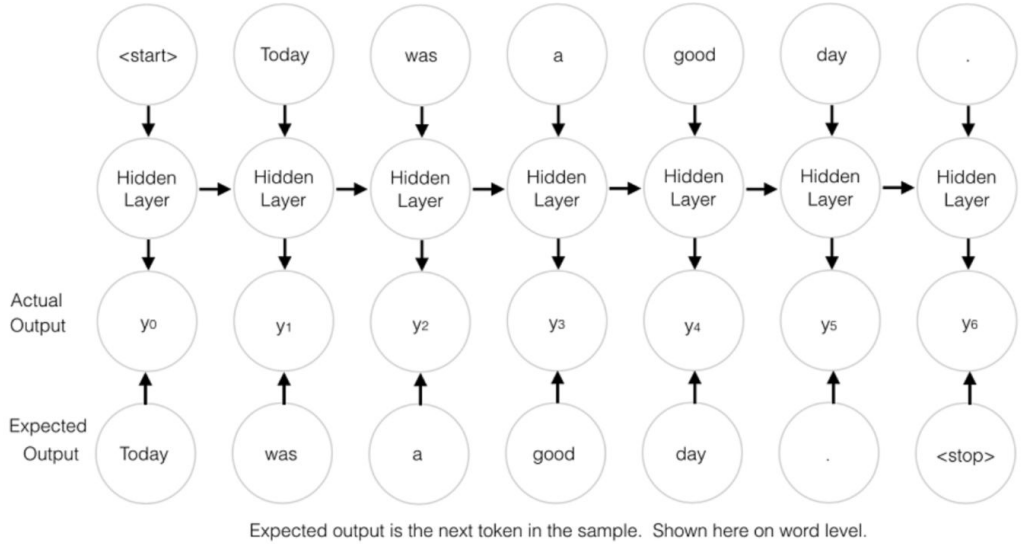


inovex

Training

- Encoder und Decoder werden *gemeinsam trainiert*
  - Trainingsdaten: Encoder-Input, Decoder-Input und Decoder-Target

- *Teacher-Forcing* für Decoder: Input und Target selbe Sequenz, aber um einen Zeitschritt verschoben



Expected output is the next token in the sample. Shown here on word level.

Hobson Lane et al.: Natural Language Processing in Action, Kap. 10, S. 375

inovex

Anwendung

1. Eingabe enkodieren
   => Thought Vector
2. Decoder mit Start-Token und Thought Vector initialisieren
   => Nächstes Token
3. Decoder dem generierten Token und aktualisiertem Thought Vector (innerer Decoder-Zustand) füttern
   => Nächstes Token
4. Schritt 3 wiederholen, bis End-Token generiert wird.

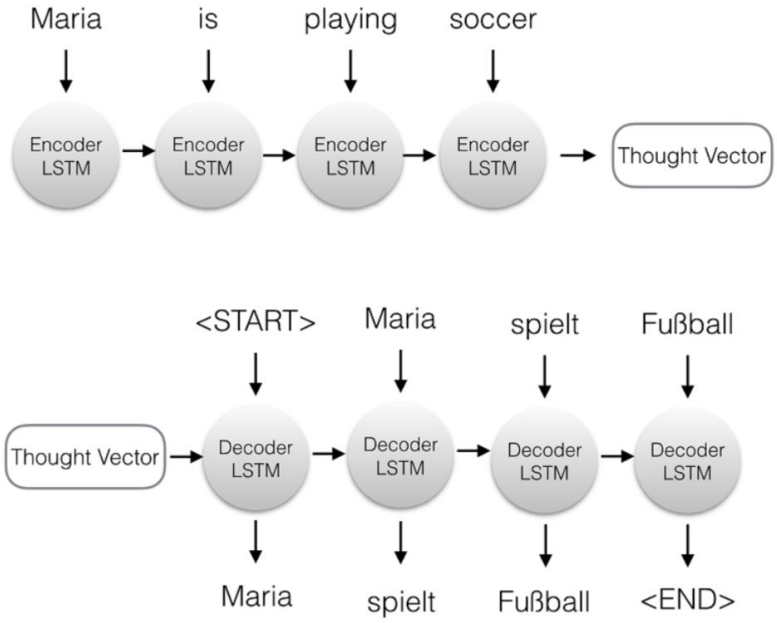


Figure 10.3 Unrolled encoder-decoder

Hobson Lane et al.: Natural Language Processing in Action, Kap. 10, S. 372

Evaluierung

- Wann ist eine Übersetzung eine gute Übersetzung?
- Wie lässt sich das quantifizieren?

Evaluierung

1. Versuch

- Quotient aus der Anzahl aller richtig übersetzten Wörtern *m* und allen Wörtern *w*
- Referenz: *the cat sat on the mat*
- Übersetzung: *the the the the the*

Evaluierung

2. Versuch

- Lösung: Begrenzung der Anzahl der richtigen Wörtern m, auf die maximale mögliche richtige Anzahl m^{ref}

$$P = \frac{\min(m, m^{ref})}{w}$$

Evaluierung

3. Versuch

- Referenz: *the cat sat on the mat*
- Übersetzung: *the the on cat sat mat*
- Lösung: Quotient aus der Anzahl aller richtig übersetzten n-Gramme m_n begrenzt auf die Anzahl der möglichen richtigen m^{ref}_n und allen n-Grammen w_n

$$P_n = \frac{\min(m_n, m_n^{ref})}{w_n}$$

Evaluierung

BLEU bilingual evaluation understudy

$$\text{BLEU} = \left(\prod_{n=1}^{4} P_n\right)^{\frac{1}{4}}$$

Evaluierung

BLEU bilingual evaluation understudy

- Problem: zu kurze Übersetzungen erhalten eine zu gute Bewertung

$$\mathrm{BLEU} = \min(1, \exp(1 - \frac{r}{c}))(\prod_{n=1}^{4} P_n)^{\frac{1}{4}}$$

Und nun …

… Spaß mit Seq2Seq-Learning

https://github.com/hskaailabnlp/machine_translation

Vielen Dank

Anna Weißhaar
aweisshaar@inovex.de

Tilman Berger
twittl@inovex.de