„Entwicklung einer Syntaxkorrektur im sprachlichen Kontext der Firma HORSCH GmbH“

Antrittsvortrag von Christoph Meyer

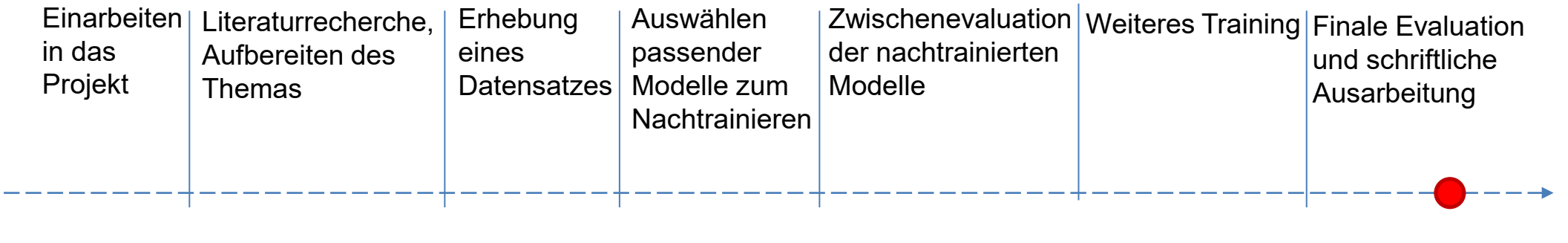
Lehrstuhl für Medieninformatik
Institut für Information und Medien, Sprache und Kultur
Fakultät für Sprach-, Literatur- und Kulturwissenschaften

Universität Regensburg

Organisatorische Daten

Semester:	7. Bachelor Semester
Medieninformatik	
Betreuer:	Prof. Dr. Christian Wolff
Erstgutachter:	-
Zweitgutachter:	-

Organisatorische Daten



Agenda

1) Hintergrund
2) Problemstellung
3) Meine Aufgabe
4) Verwandte Arbeiten
5) Mein Ansatz
6) Weiteres Vorgehen

Hintergrund

#01

Background – Aufbau einer Wissensdatenbank der Firma HORSCH

https://www.smart-zebra.de/wp-content/uploads/2020/02/smartZebra-wissensdatenbank-05-300x300.jpg

Background – Weltweite Vernetzung von Kunden/Händlern/etc.

https://www.daimler.com/nachhaltigkeit/grundlagen/beschaeftigte/netzwerke.html

Problemstellung

#02

Problemstellung

- Problem: Sprachliche Barriere zwischen Menschen aus unterschiedlichen Teilen der Welt

→ Ausgangsbasis: Englisch als Weltsprache

https://blog.manager-institut.de/stolperfalle-interkulturelle-kommunikation/

Meine Aufgabe

#03

Meine Aufgabe

- Entwicklung einer Syntaxkorrektur für diese Plattform
- Umsetzung via Deep Learning

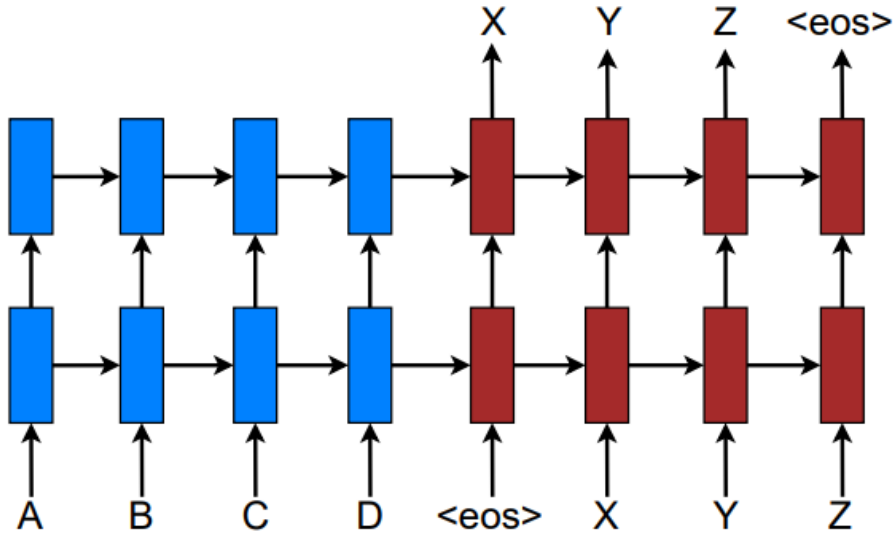
→ Nachtrainieren und Vergleich bestehender Modelle

Verwandte Arbeiten

#04

Related Work – Neural Machine Translation (NMT)

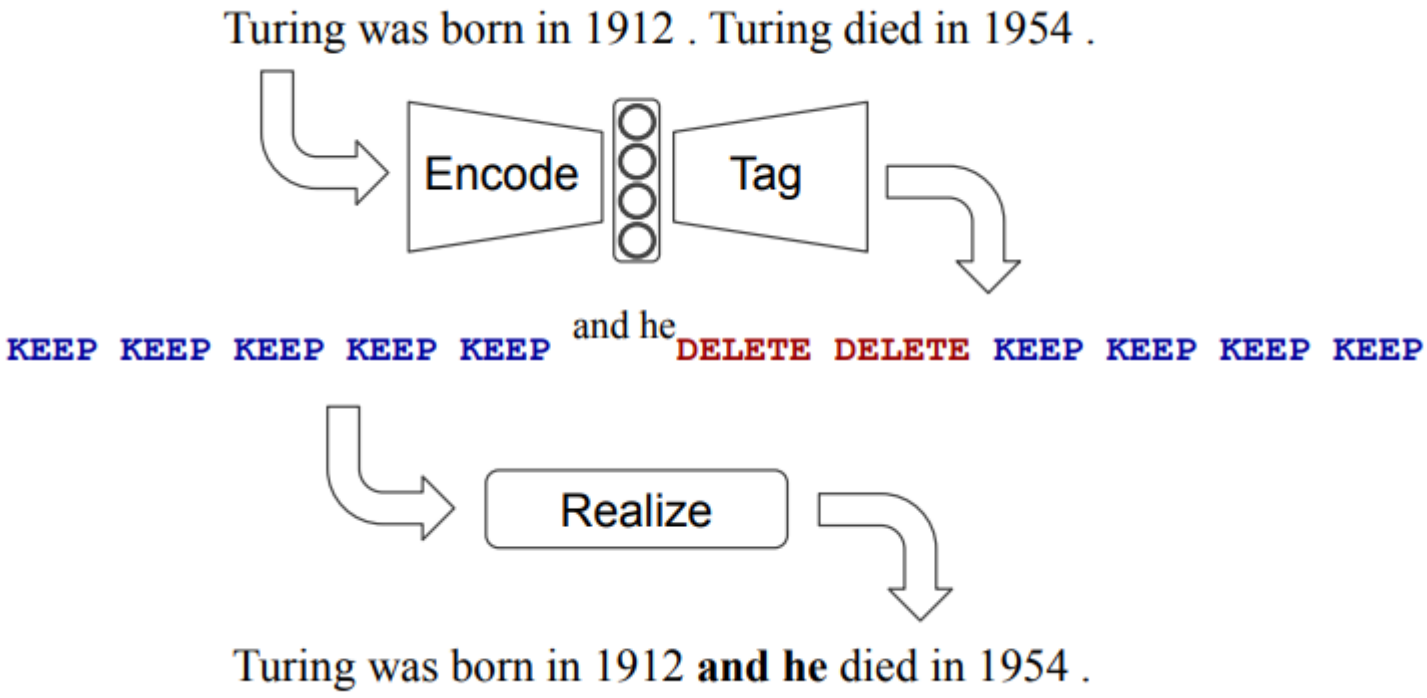
- RNN-Encoder liest Satz Token für Token ein
- Dadurch entsteht ein State-Vektor mit fester Größe
- RNN-Decoder bildet Tokenweise den Zielsatz anhand des State-Vektors
- Mögliche Ergänzungen: Attentionmechanismus, Residual Connections



Luong et al. (2015): Effective Approaches to Attention-based Neural Machine Translation

Vaswani et al. (2018): Tensor2Tensor for Neural Machine Translation

Related Work – Sequence Tagging



Malmi et al. (2019): Encode, Tag, Realize: High-Precision Text Editing

Mein Ansatz

#05

Grammatical Error Correction: Tag, Not Rewrite

- GEC-Modell, das als Basis verschiedene Transformer Modelle verwendet
- Ziel: Fehler annotieren anstatt Text zu generieren
- Erreicht durch Klassifizieren der Tokens über alle Fehlerklassen

Omelianchuk et al. (2020): GECToR – Grammatical Error Correction: Tag, Not Rewrite

Parallel Iterative Edit Models for Sequence Transduction

- GEC-Modell auf BERT-Basis
- Verfolgt ebenfalls einen Sequence-Labeling-Ansatz
- Durch paralleles Vorgehen und iterative Verbesserungen kann die Performance weiter optimiert werden

Awasthi et al. (2019): Parallel Iterative Edit Models for Local Sequence Transduction

Improving Grammatical Error Correction via Pre-Training a Copy Augmented Architecture with Unlabeled Data

- GEC-Modell, das es erlaubt korrekte Tokens/Sequenzen in den Zielsatz zu kopieren

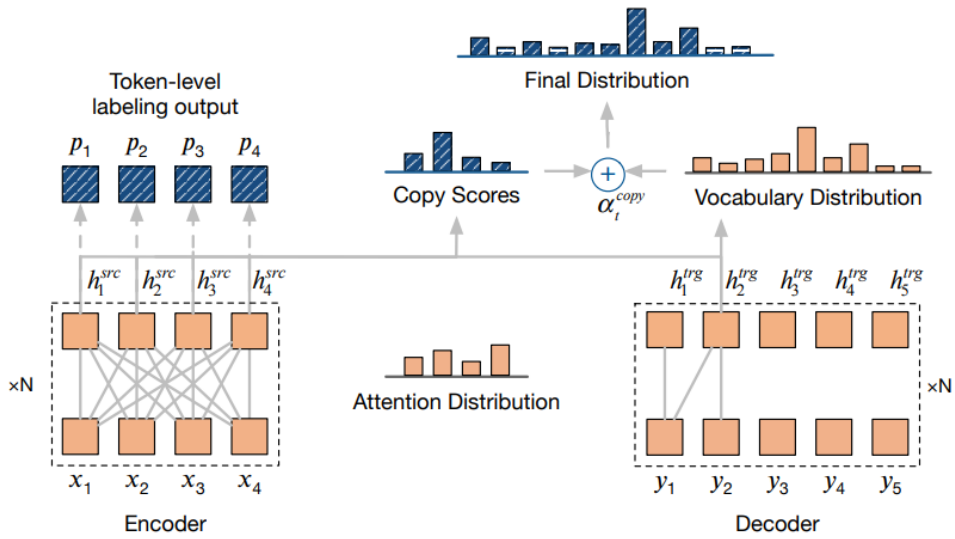


Figure 1: Copy-Augmented Architecture.

Zhao et al. (2019): Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data

Ergebnisse der Zwischenevaluation

Architektur	Modell	F0.5-Score	Steigerung
GECToR	BERT (ursprüngliches Modell)	55,31	+8,93
	BERT (nachtrainiertes Modell)	64,24	
	RoBERTa (ursprüngliches Modell)	65,57	+4,34
	RoBERTa (nachtrainiertes Modell)	69,91	
	XLNet (ursprüngliches Modell)	63,91	+6,02
	XLNet (nachtrainiertes Modell)	**69,93**	
Fairseq-GEC	Ursprüngliches Modell	29,35	**+24,08**
	Nachtrainiertes Modell	53,43	

Weiteres Vorgehen

#06

Weiteres Vorgehen

- Erneutes Nachtrainieren bis insgesamt 20 Epochen erreicht sind
- Auswahl des besten Modells aus den 20 Epochen
- Erneute Evaluation
- Schriftliche Ausarbeitung vervollständigen

„Entwicklung einer Syntaxkorrektur im sprachlichen Kontext der Firma HORSCH GmbH“

Lehrstuhl für Medieninformatik
Institut für Information und Medien, Sprache und Kultur
Fakultät für Sprach-, Literatur- und Kulturwissenschaften