Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

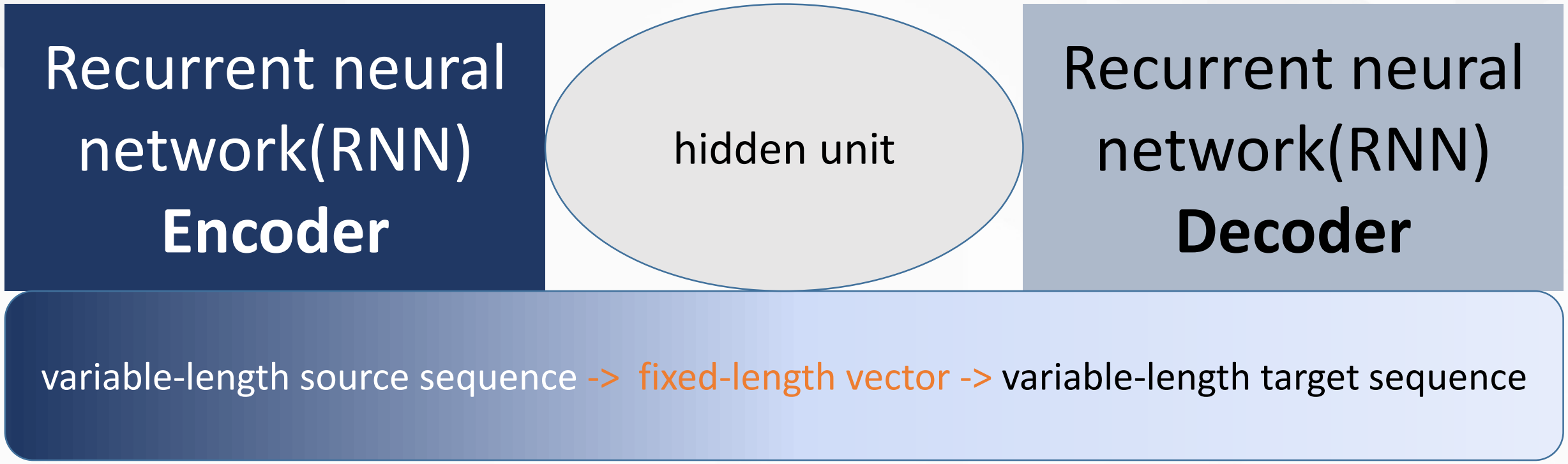
Yupeng Guo

Agenda

- Introduction
- RNN Encoder-Decoder
  - Recurrent Neural Networks
  - RNN Encoder–Decoder
  - Hidden Unit that Adaptively Remembers and Forgets
- Statistical Machine Translation
  - Definition and examples of SMT
  - Scoring Phrase Pairs with RNN Encoder–Decoder
- Experiments
  - Data and Baseline System
  - Quantitative Analysis
  - Qualitative Analysis
  - Word and Phrase Representations
- Conclusion & Outlook

Einleitung

1. Von Deep neural networks zu SMT (Statistical Machine Translation).
2. RNN Encoder–Decoder in Phrase-based SMT system.



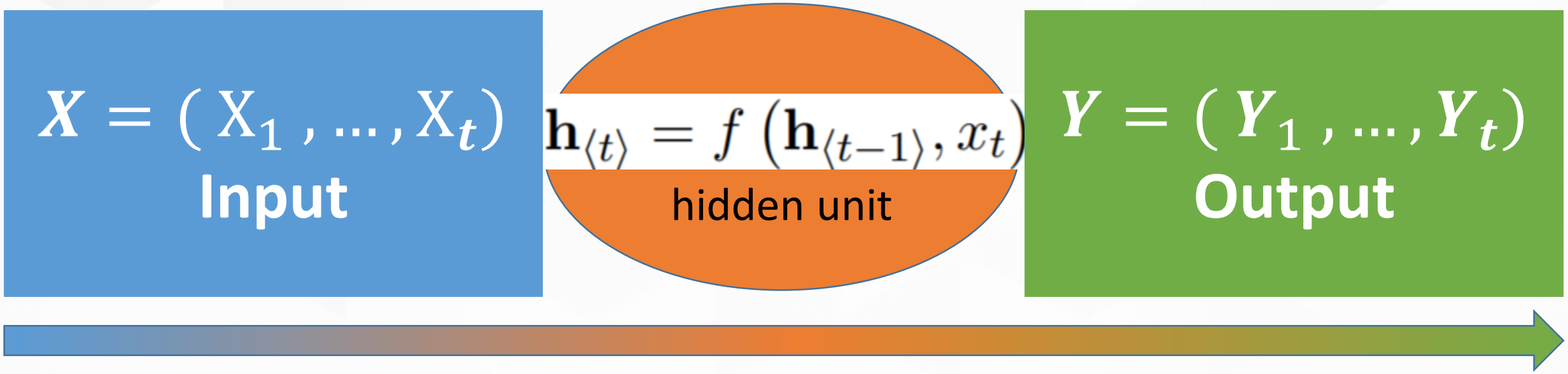
Agenda

- Introduction
- RNN Encoder-Decoder
  - Recurrent Neural Networks
  - RNN Encoder–Decoder
  - Hidden Unit that Adaptively Remembers and Forgets
- Statistical Machine Translation
  - Definition and examples of SMT
  - Scoring Phrase Pairs with RNN Encoder–Decoder
- Experiments
  - Data and Baseline System
  - Quantitative Analysis
  - Qualitative Analysis
  - Word and Phrase Representations
- Conclusion & Outlook

RNN Encoder-Decoder

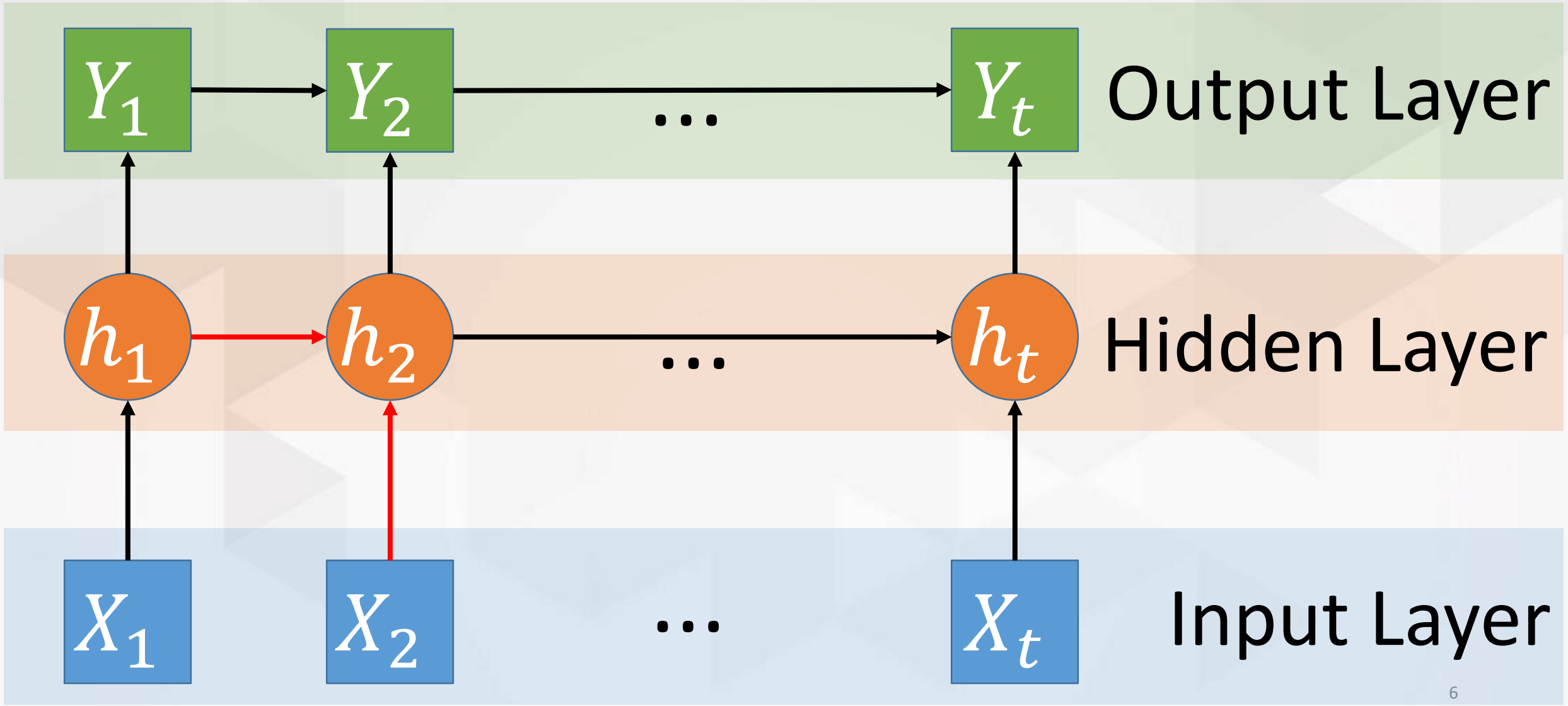
Vorläufig: Recurrent Neural Networks

Ein RNN ist ein neuronales Netzwerk, das aus einem **“hidden state”** h und einem optionalen **Ausgang** y besteht, der auf einer “variable-length sequence **Eingang”** x = (x1,..., XT) operiert.



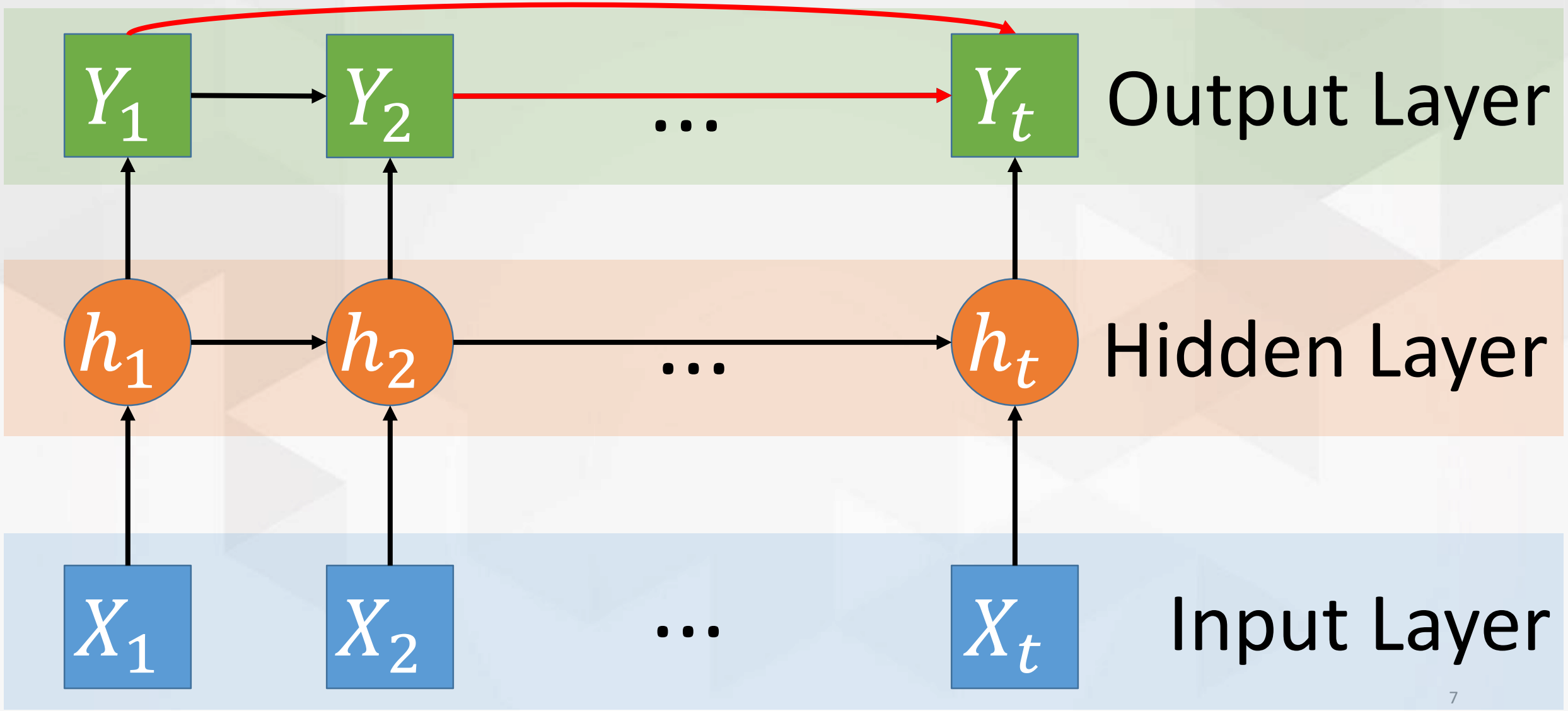
Recurrent Neural Networks

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right)$$

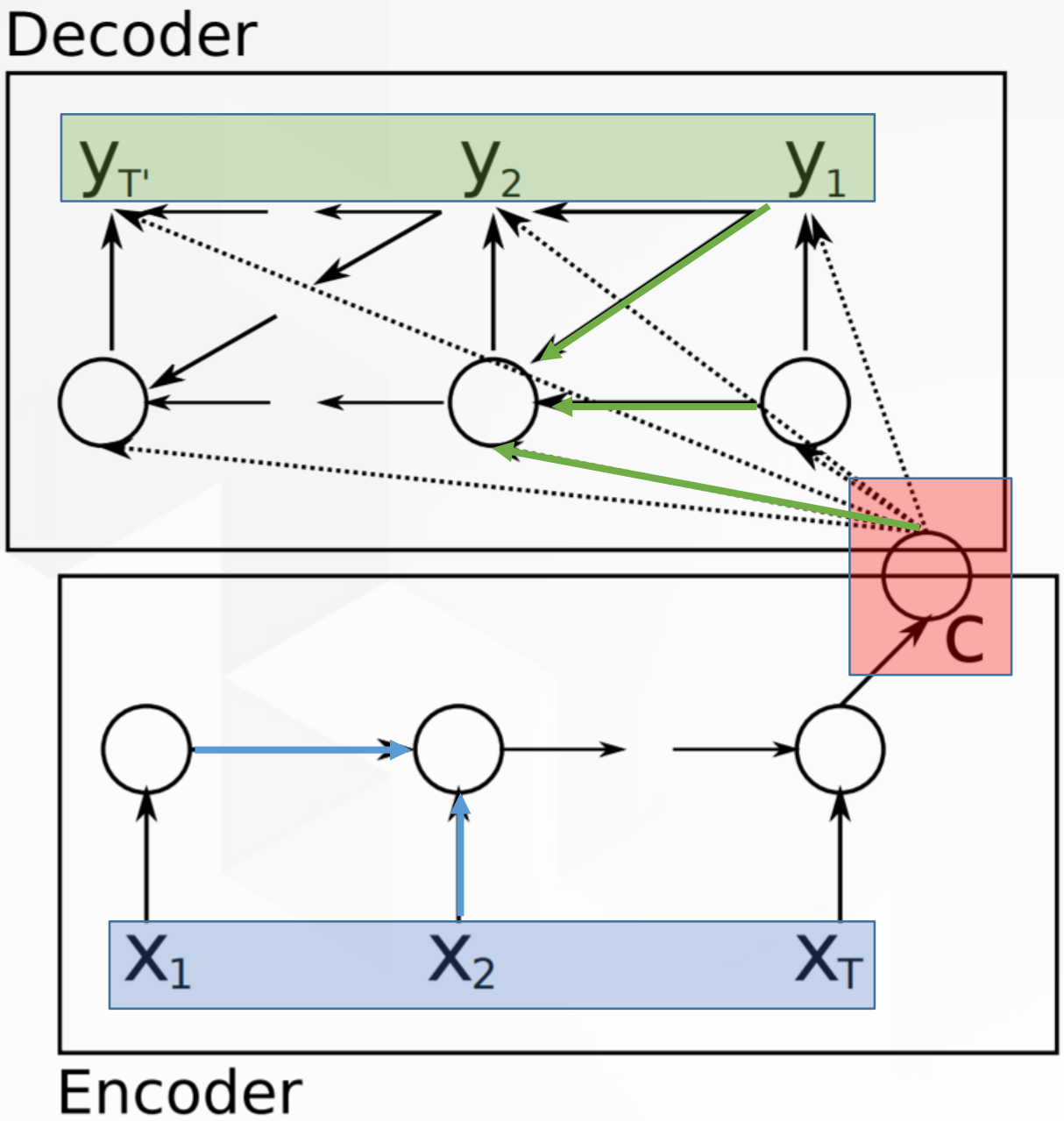


Recurrent Neural Networks

$$p(y_t) = p(y_t \mid y_{t-1}, \ldots, y_1)$$



RNN Encoder-Decoder



$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

Gl.(2)

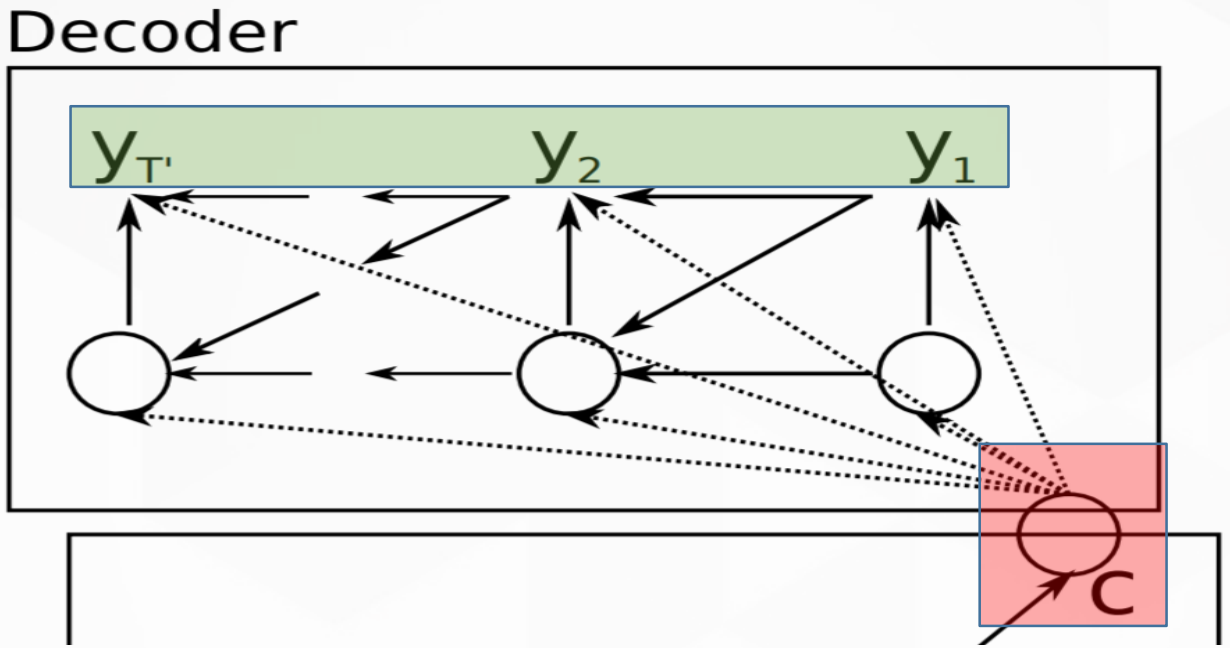
$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right)$$

Gl.(1)

RNN Encoder-Decoder

Und ähnlich, die bedingte Verteilung des nächsten Symbols ist

$$P(y_t|y_{t-1}, y_{t-2}, \cdots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right)$$



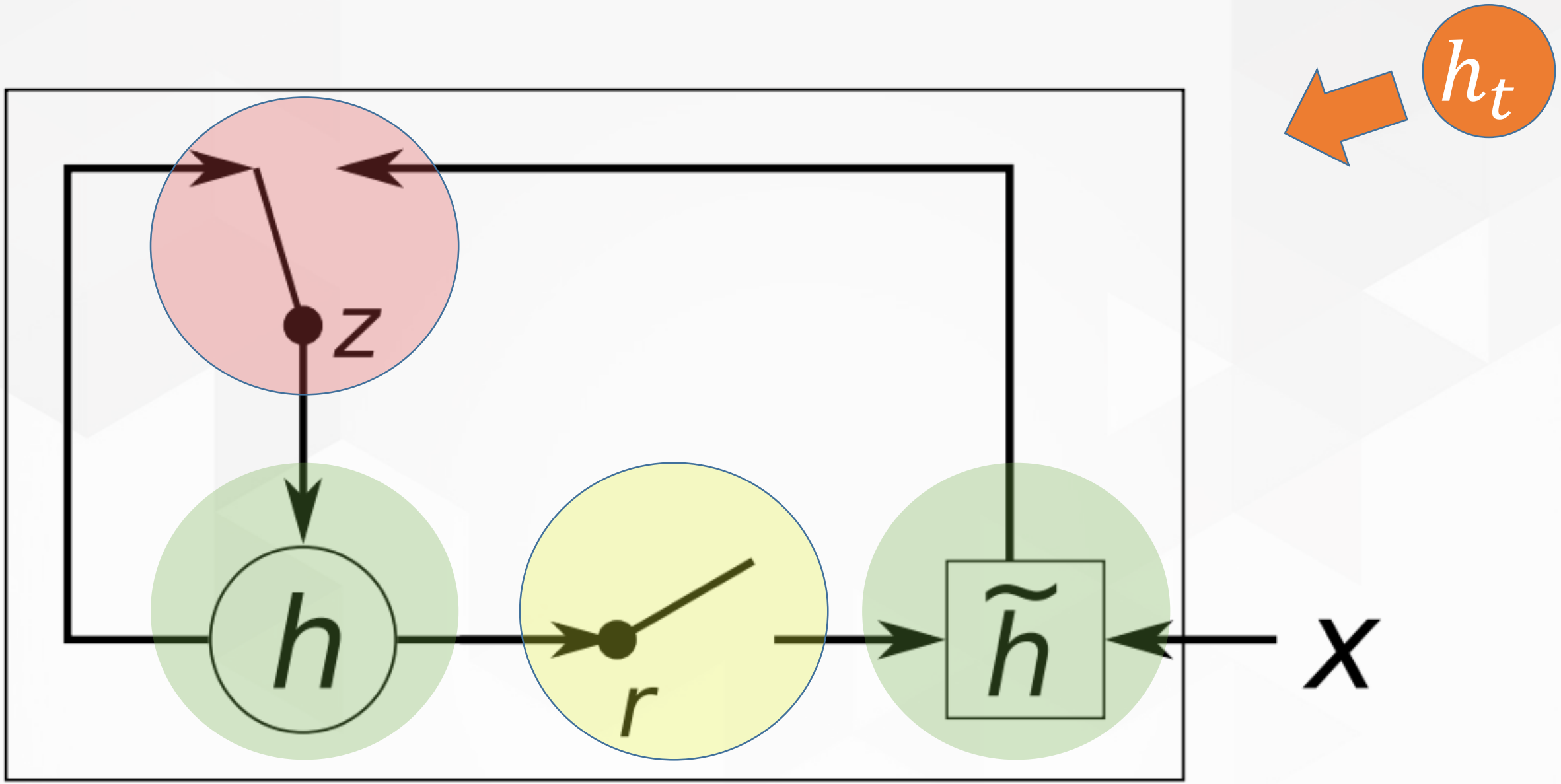
RNN Encoder-Decoder

Sobald der RNN Encoder-Decoder trainiert ist, kann das Modell auf zwei Arten verwendet werden.

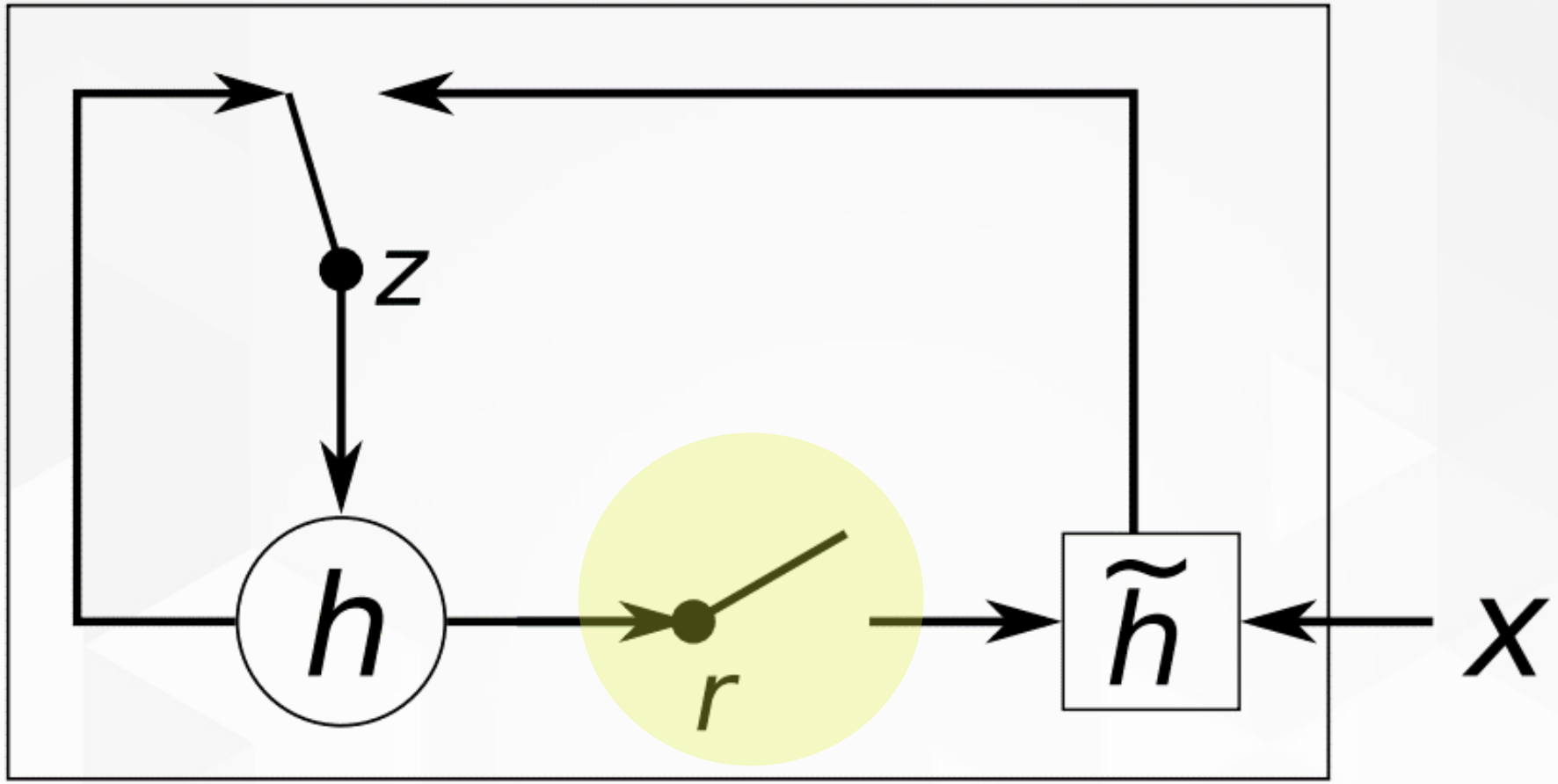
1. Das Modell kann verwendet werden, um eine Zielsequenz bei einer gegebenen Eingabesequenz zu erzeugen.
2. Das Modell kann verwendet werden, um ein gegebenes Paar von Eingabe- und Ausgabesequenzen zu bewerten, wobei die Bewertung einfach eine Wahrscheinlichkeit pθ ist.

$$\log p_{\boldsymbol{\theta}}(\mathbf{y}_n \mid \mathbf{x}_n)$$

Hidden Unit that Adaptively Remembers and Forgets

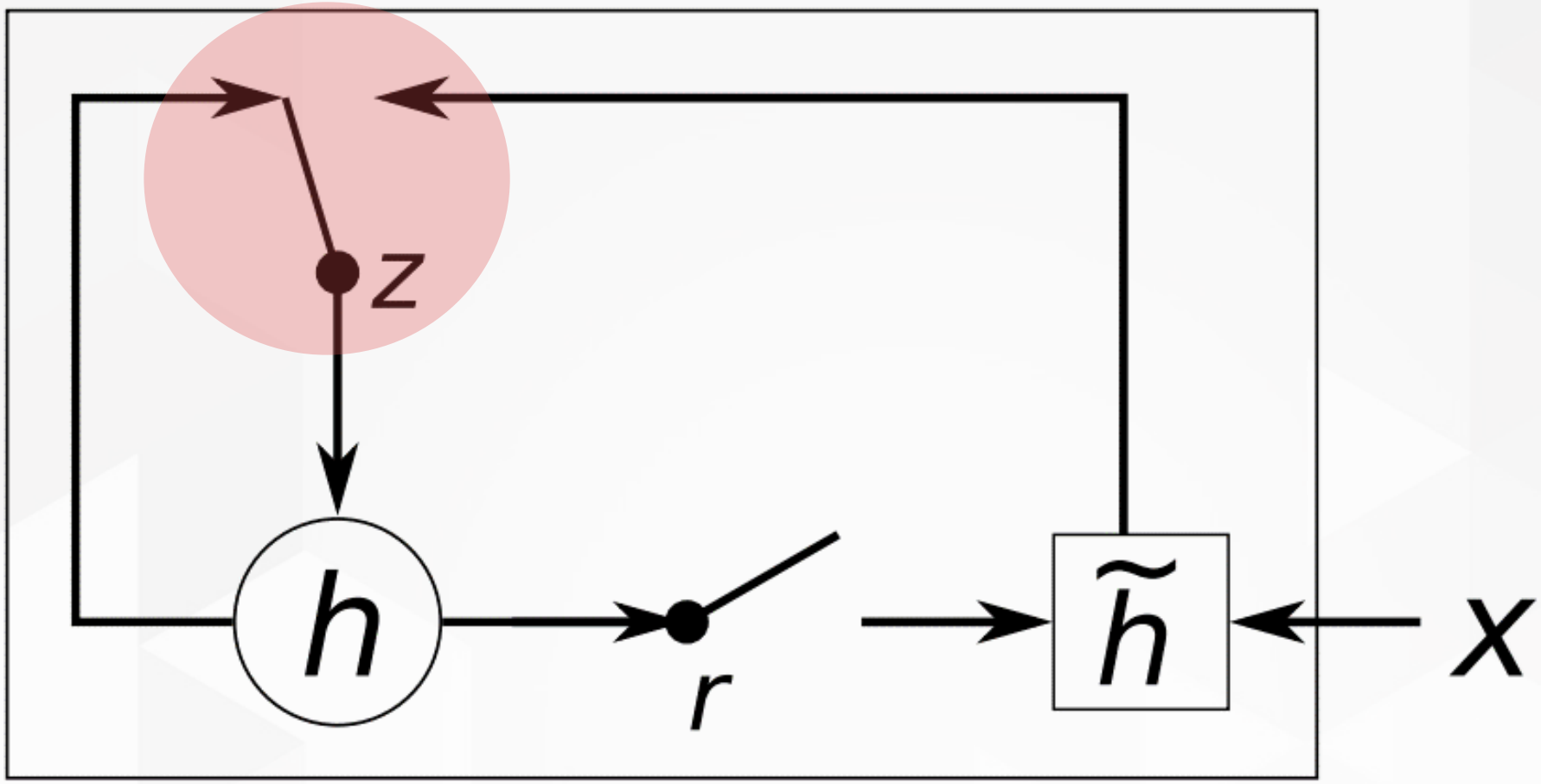


Hidden Unit that Adaptively Remembers and Forgets



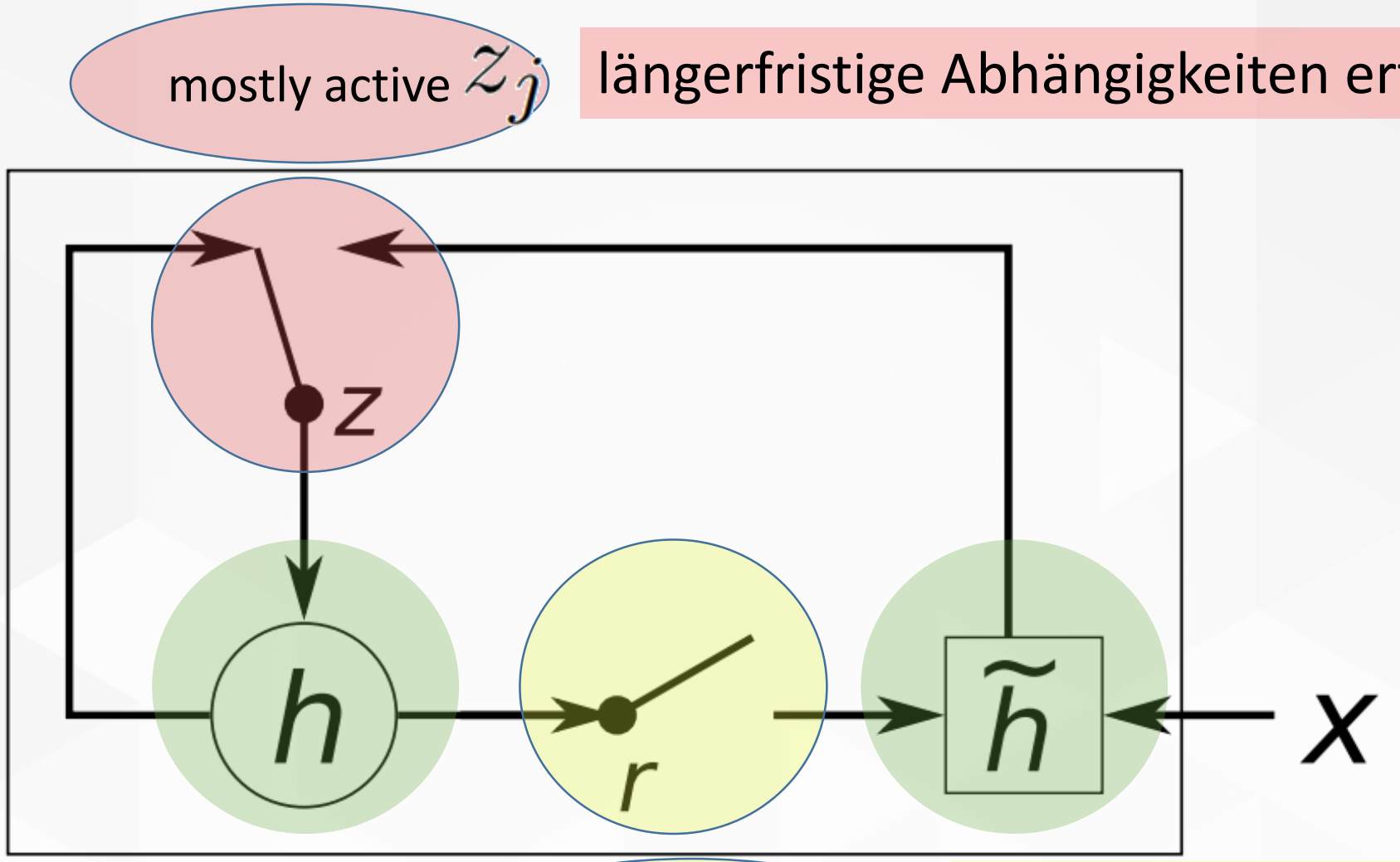
$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

Hidden Unit that Adaptively Remembers and Forgets



$$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

Hidden Unit that Adaptively Remembers and Forgets



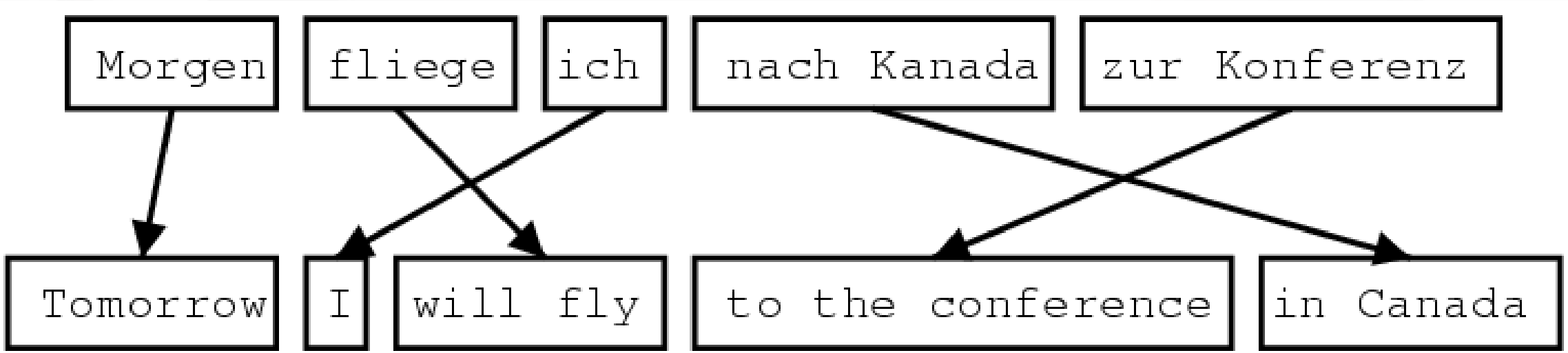
Agenda

- Introduction
- RNN Encoder-Decoder
  - Recurrent Neural Networks
  - RNN Encoder–Decoder
  - Hidden Unit that Adaptively Remembers and Forgets
- Statistical Machine Translation
  - Definition and examples of SMT
  - Scoring Phrase Pairs with RNN Encoder–Decoder
- Experiments
  - Data and Baseline System
  - Quantitative Analysis
  - Qualitative Analysis
  - Word and Phrase Representations
- Conclusion & Outlook

Statistical Machine Translation

- Statistische Analyse des Parallelkorpus
- Konstruieren des statistischen Übersetzungsmodells
- Wort, Phrase, Syntax - basierte Übersetzung

Bsp. Phrase-based SMT (alignment)



Statistical Machine Translation

$$p(\mathbf{f} \mid \mathbf{e}) \propto p(\mathbf{e} \mid \mathbf{f})p(\mathbf{f})$$

f → Übersetzungen
e → Quellsatz

Übersetzungswahrscheinlichkeiten für “matching” Phrasen in den Quell- und Zielsätzen

Translation model, (Koehn et al., 2003) (Marcu and Wong, 2002)

Eine Wahrscheinlichkeitsverteilung über Sequenzen von Wörtern

Language model, (Koehn, 2005)

Ziel: Übersetzungsqualität zu maximieren

Scoring Phrase Pairs with RNN Encoder–Decoder

train the RNN Encoder–Decoder

- an einer Tabelle von Phrasenpaaren trainiert wurde
- Die (normalisierten) Frequenzen jedes Phrasenpaars in den ursprünglichen Korpora wurden ignoriert.
- Sobald der RNN-Encoder-Decoder trainiert ist, wird eine neue Bewertung für jedes Phrasenpaar zu der existierenden Phrasentabelle hinzugefügt.

Agenda

- Introduction
- RNN Encoder-Decoder
  - Recurrent Neural Networks
  - RNN Encoder–Decoder
  - Hidden Unit that Adaptively Remembers and Forgets
- Statistical Machine Translation
  - Definition and examples of SMT
  - Scoring Phrase Pairs with RNN Encoder–Decoder
- Experiments
  - Data and Baseline System
  - Quantitative Analysis
  - Qualitative Analysis
  - Word and Phrase Representations
- Conclusion & Outlook

Experimente

Der Ansatz wurde auf der Englisch / Französisch-Übersetzungsaufgabe des WMT'14-Workshops evaluiert.

- zweisprachiges Korpora.
- Europarl (61 Millionen Wörter)
- Nachrichtenkommentare (5,5 Millionen)
- UN (421 Millionen)
- zwei „crawled“ Korpora von 90 Millionen Wörtern und 780 Millionen Wörtern.

Data and Baseline System

Alle Daten → Schlechte Leistung, Sehr großes Modell

Lösungen (Datenauswahlverfahren): (Moore and Lewis, 2010), (Axelrod et al., 2011)

ein " baseline phrase-based SMT system " etablieren :

Sprachmodellierung:

Trainieren des RNN-En-Des:

Set für Datenauswahl ,Gewichtsabstimmung und Testset:

(Jeder Set hat mehr als 70.000 Wörter und eine einzige Referenzübersetzung)

Data and Baseline System

Datenauswahl in Training

Zum Training der neuronalen Netze und RNN-Encoder-Decoder.

- Die Quelle war begrenzt und zielte auf Vokabeln zu den häufigsten 15.000 Wörtern für Englisch und Französisch. Dies deckt ungefähr 93% des Datensatzes ab.

Data and Baseline System

RNN Encoder-Decoder in Training

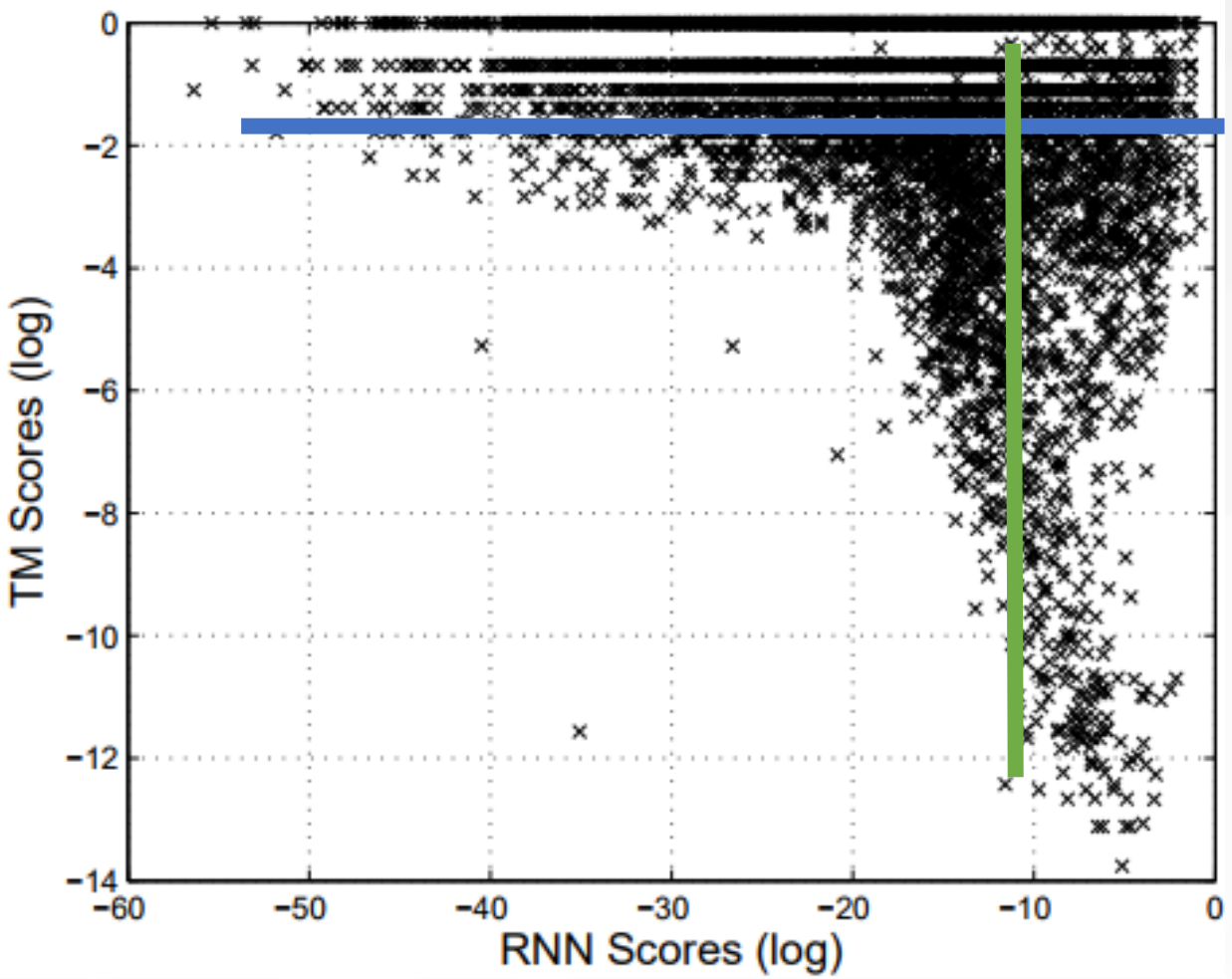
- 1000 versteckte Einheiten
- Die Eingabe-/Ausgangsmatrix (zwischen jedem Eingabesymbol $X_{<t>}$ und hidden unit) wird mit zwei lower-rank Matrizen approximiert.
- Rank-100 -Matrizen wurden verwendet, entspricht einer Einbettung der Dimension 100 für jedes Wort.
- Bei jedem Update wurden 64 zufällig ausgewählte Phrasenpaare aus einer Phrasentabelle (die aus 348 Millionen Wörtern erstellt wurde) verwendet. Das Modell wurde für ca. 3 Tage trainiert.

Data and Baseline System

Um die Effektivität zu beurteilen →

traditioneller Ansatz,
the SMT system using CSLM
(näml. TM) (Schwenk, 2007)

der vorgeschlagene Ansatz,
Scoring-Phrasen-Paaren durch
RNN Encoder-Decoder



Der Vergleich wird klären, ob sich die Beiträge von mehreren neuronalen Netzen in verschiedenen Teilen des SMT-Systems addieren oder redundant sind.

Quantitative Analysis

Kombinationen:

1. Baseline configuration
2. Baseline + RNN
3. Baseline + CSLM + RNN

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64

Addieren,
Nicht Redundant

Qualitative Analysis

Woher die Leistungsverbesserung kommt?

· Erwartet:
- bessere Scores für die häufigen Phrasen
- schlechte Scores für die seltenen Phrasen

· Weitere erwartet:
- ohne Frequenzinformation trainiert wurde

· Also achten wir auf:
- Die Paare, deren Quellphrase lang ist (mehr als 3 Wörter pro Quellphrase) und häufig ist.
- Die Paare, deren Quellphrase im Korpus lang, aber selten ist.

Qualitative Analysis

Source	Translation Model	RNN Encoder–Decoder
at the end of the	[a la fin de la] [ŕ la fin des années] [être supprimés à la fin de la]	[à la fin du] [à la fin des] [à la fin de la]
for the first time	[**r** ⓒ pour la premi**r**ëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]	[pour la première fois] [pour la première fois ,] [pour la première fois que]
in the United States and	[**?** aux **?**tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]	[aux Etats-Unis et] [des Etats-Unis et] [des États-Unis et]
, as well as	[**?**s , qu'] [**?**s , ainsi que] [**?**re aussi bien que]	[, ainsi qu'] [, ainsi que] [, ainsi que les]
one of the most	[**?**t **?**l' un des plus] [**?**l' un des plus] [être retenue comme un de ses plus]	[l' un des] [le] [un des]

(a) Long, frequent source phrases

parts of the world .	[ⓒ gions du monde .] [régions du monde considérées .] [région du monde considérée .]	[parties du monde .] [les parties du monde .] [des parties du monde .]
the past few days .	[le petit texte .] [cours des tout derniers jours .] [les tout derniers jours .]	[ces derniers jours .] [les derniers jours .] [cours des derniers jours .]
on Friday and Saturday	[vendredi et samedi à la] [vendredi et samedi à] [se déroulera vendredi et samedi ,]	[le vendredi et le samedi] [le vendredi et samedi] [vendredi et samedi]

(b) Long, rare source phrases

Qualitative Analysis

Source	Samples from RNN Encoder–Decoder
at the end of the	[à la fin de la] (×11)
for the first time	[pour la première fois] (×24) [pour la première fois que] (×2)
in the United States and	[aux États-Unis et] (×6) [dans les États-Unis et] (×4)
, as well as	[, ainsi que] [,] [ainsi que] [, ainsi qu'] [et UNK]
one of the most	[l' un des plus] (×9) [l' un des] (×5) [l' une des plus] (×2)

(a) Long, frequent source phrases

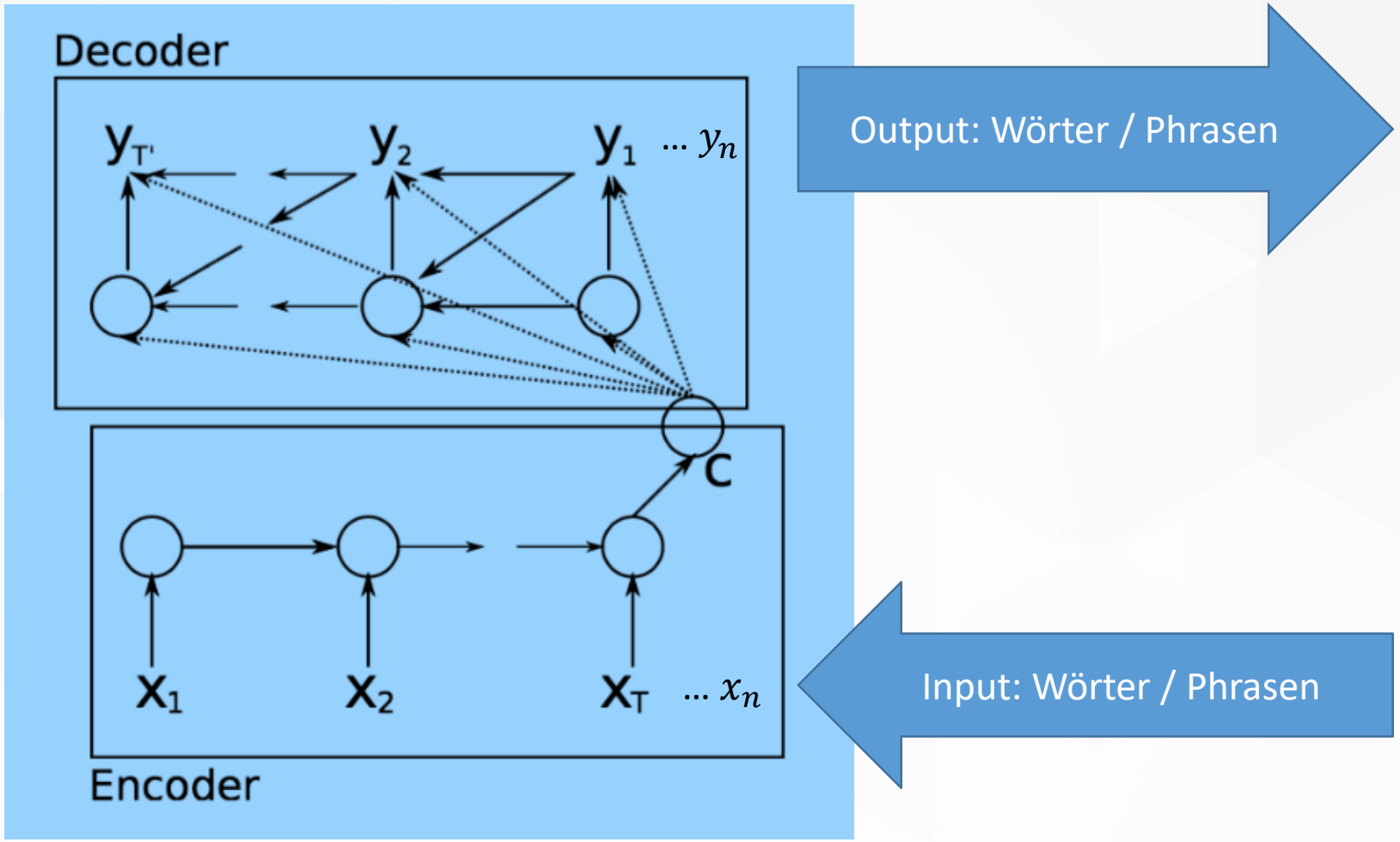
Source	Samples from RNN Encoder–Decoder
, Minister of Communications and Transport	[, ministre des communications et le transport] (×13)
did not comply with the	[n' tait pas conforme aux] [n' a pas respect l'] (×2) [n' a pas respect la] (×3)
parts of the world .	[arts du monde .] (×11) [des arts du monde .] (×7)
the past few days .	[quelques jours .] (×5) [les derniers jours .] (×5) [ces derniers jours .] (×2)
on Friday and Saturday	[vendredi et samedi] (×5) [le vendredi et samedi] (×7) [le vendredi et le samedi] (×4)

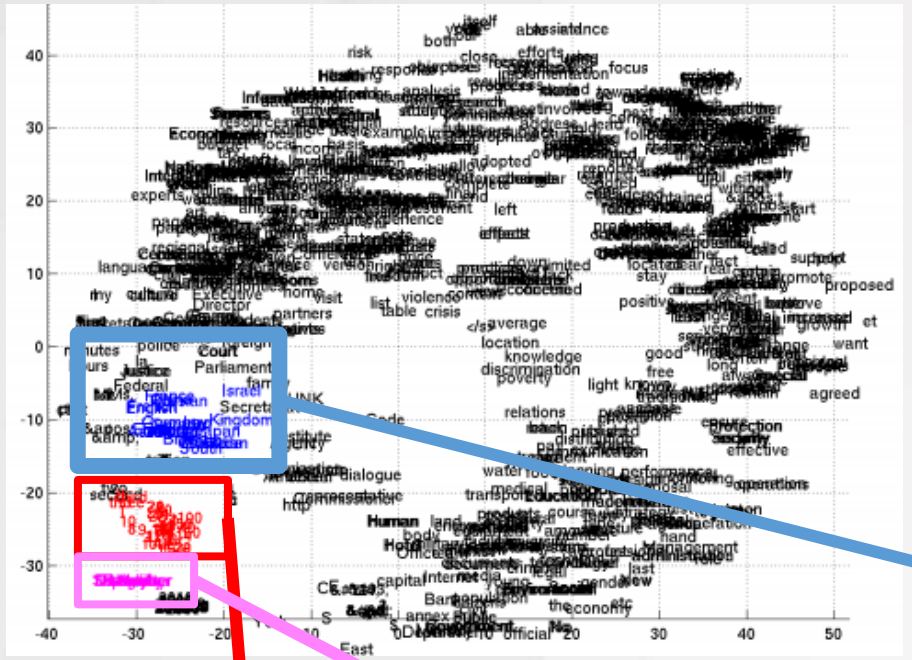
(b) Long, rare source phrases

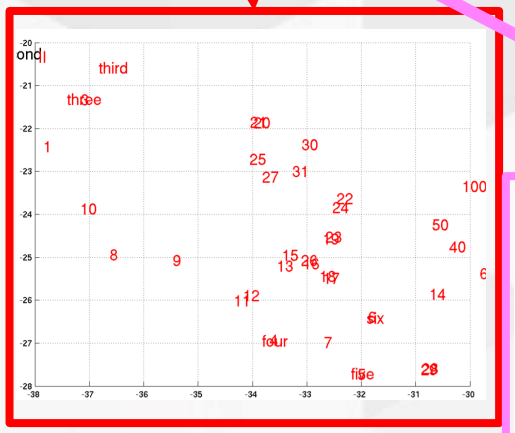
RNN Encoder-Decoder ersetzen die ganze oder einen Teil der Phrasentabelle des „standard phrase-based SMT system" in der Zukunft?

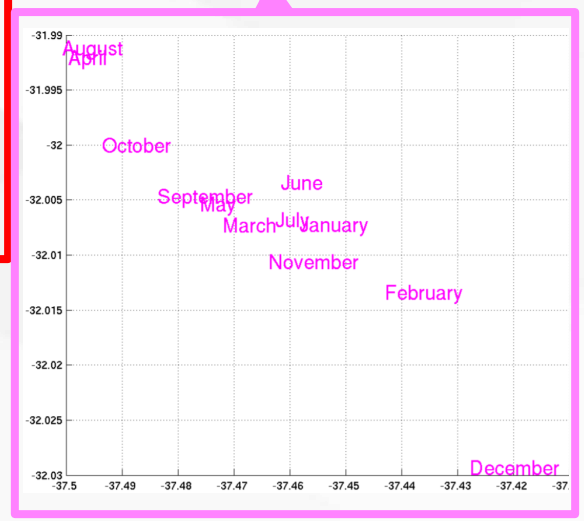
Wort und Phrase Repräsentationen

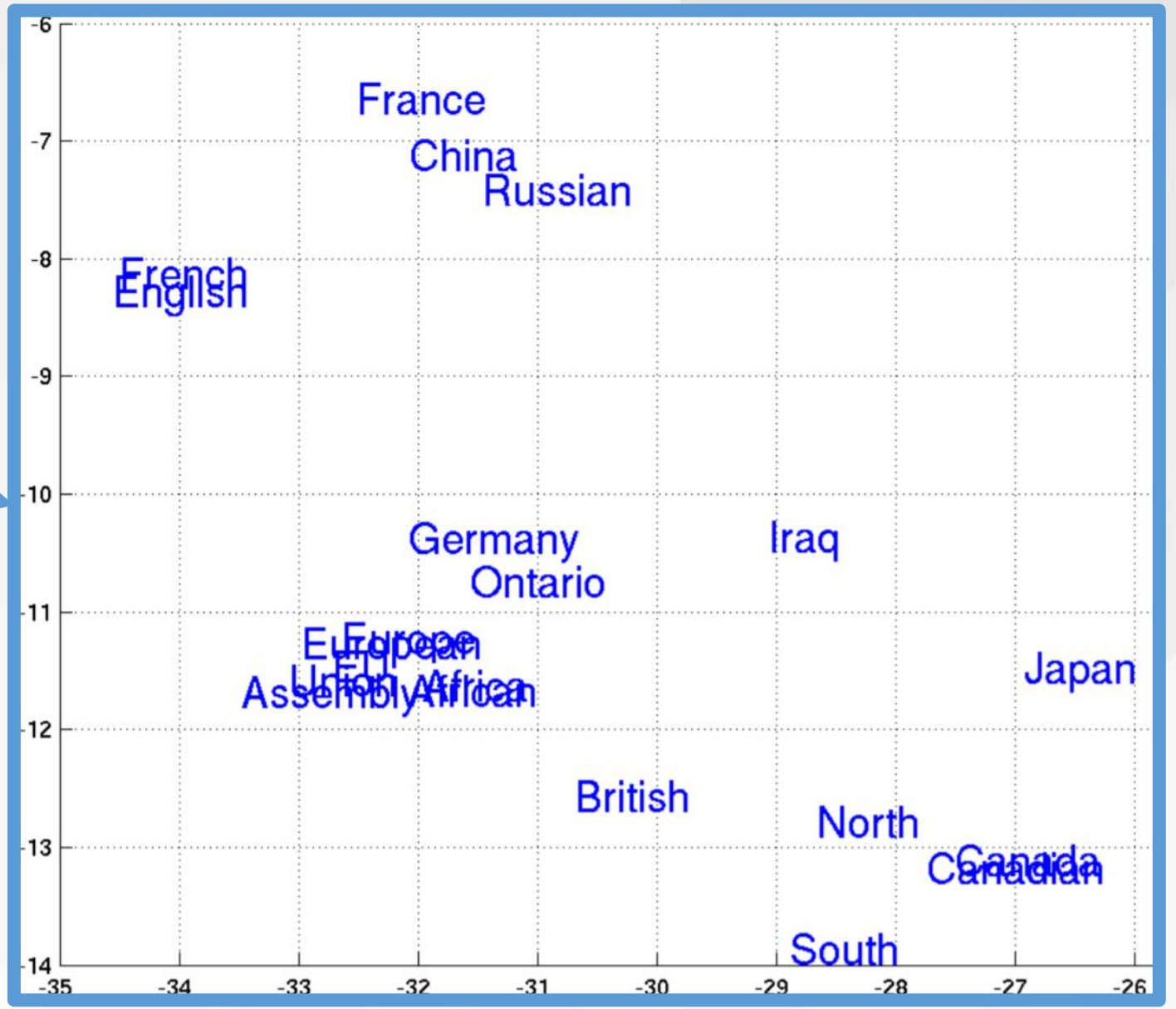
RNN Encoder-Decoder projiziert eine Folge von Wörtern in einen kontinuierlichen Raumvektor und bildet sie dann zurück.



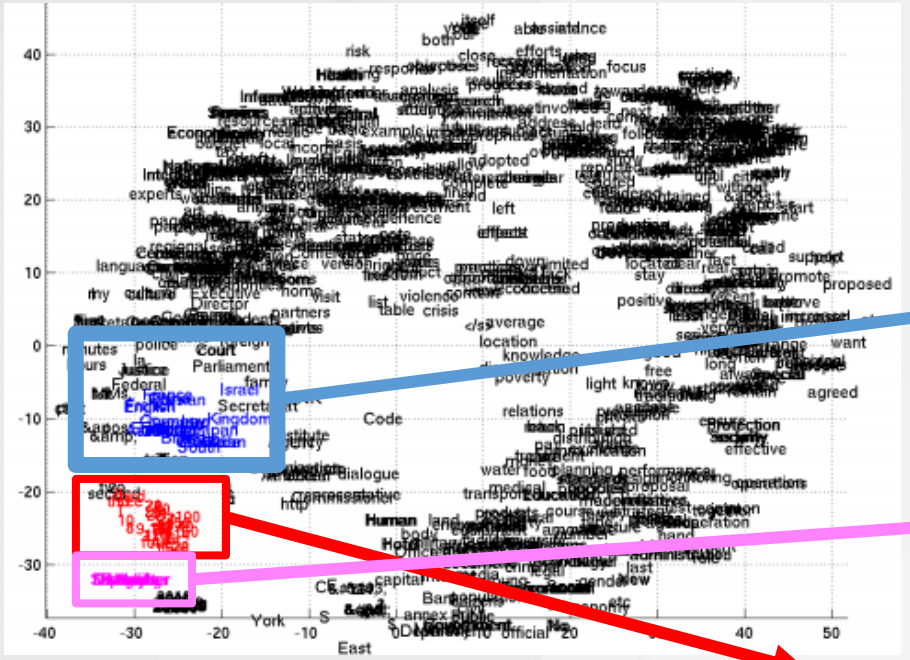


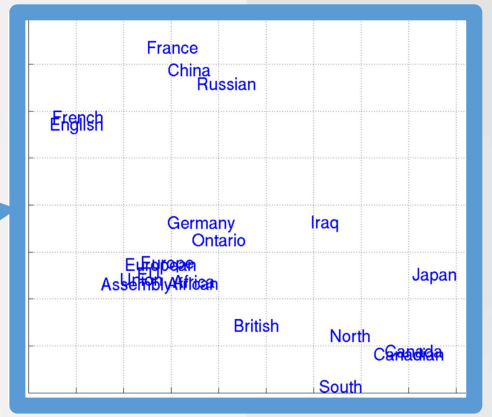


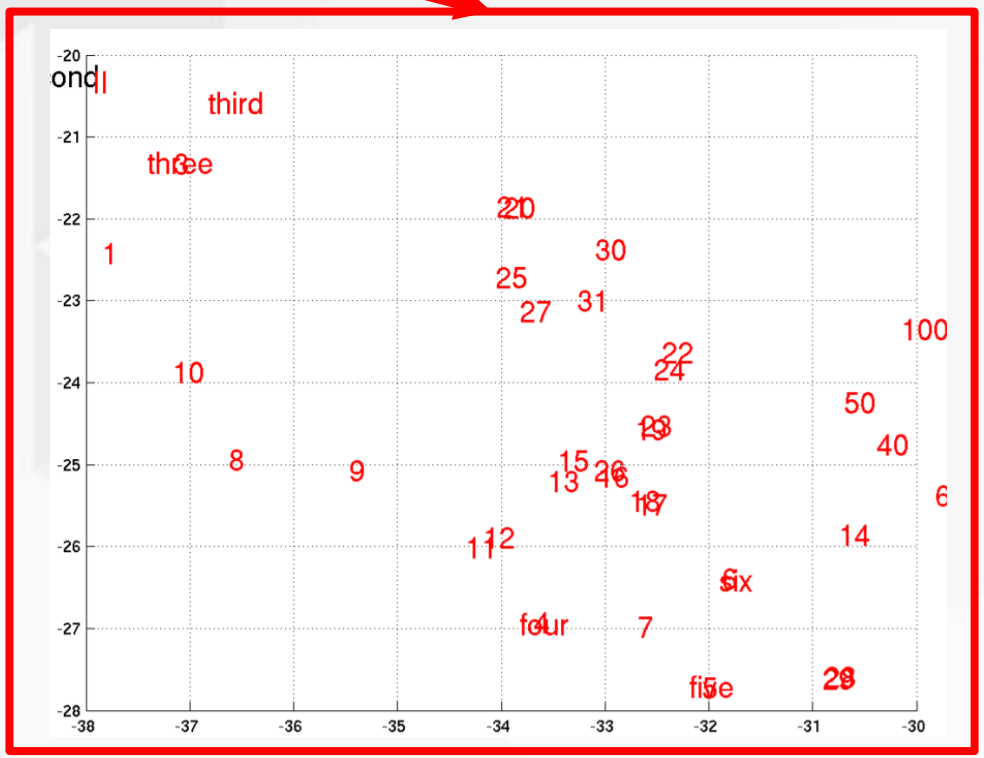


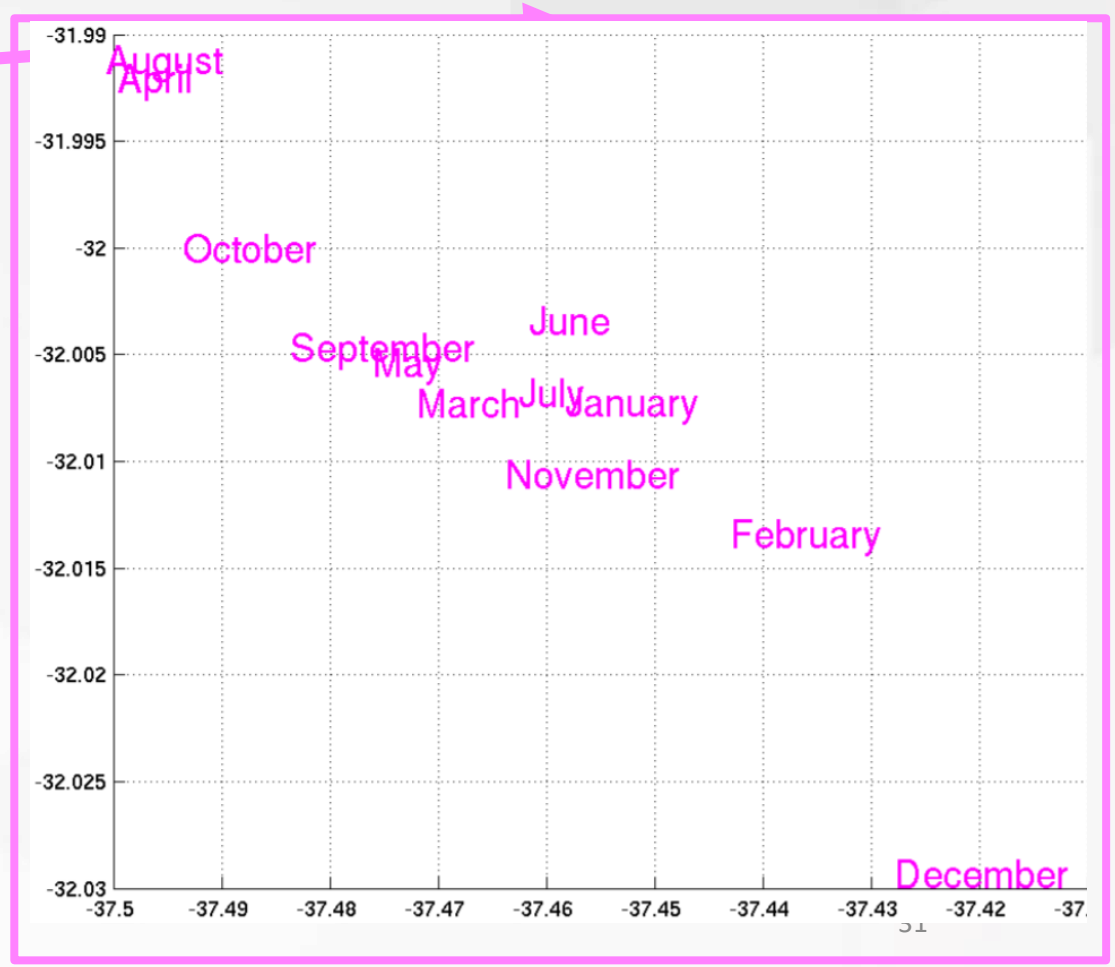


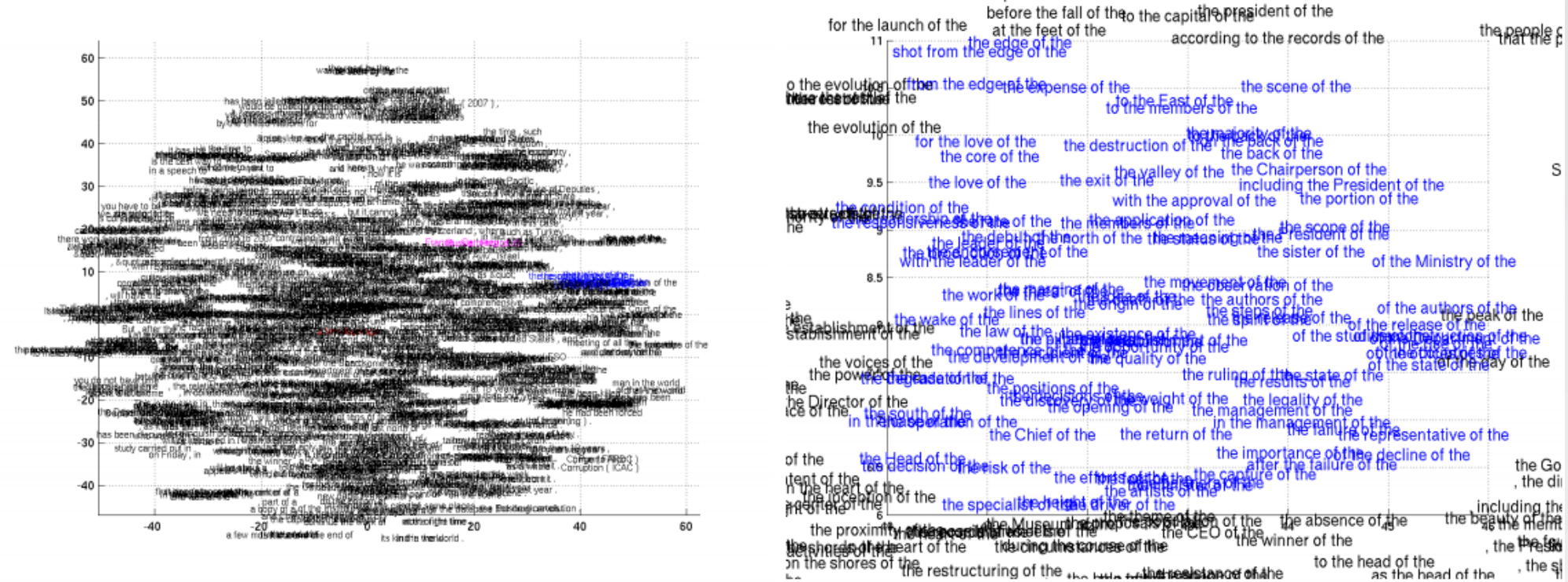
Wort Repräsentationen

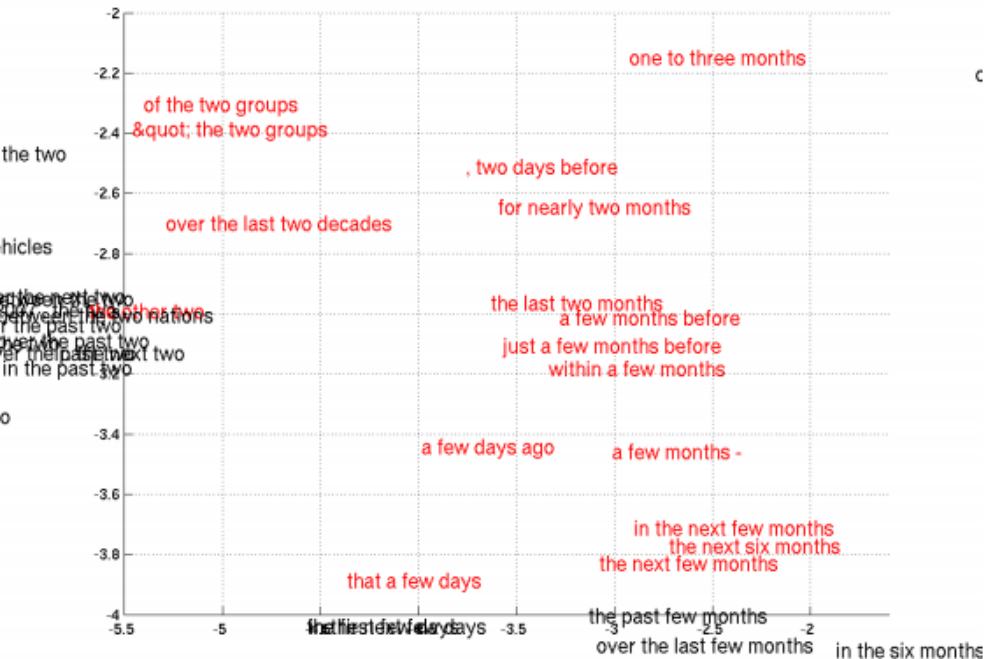


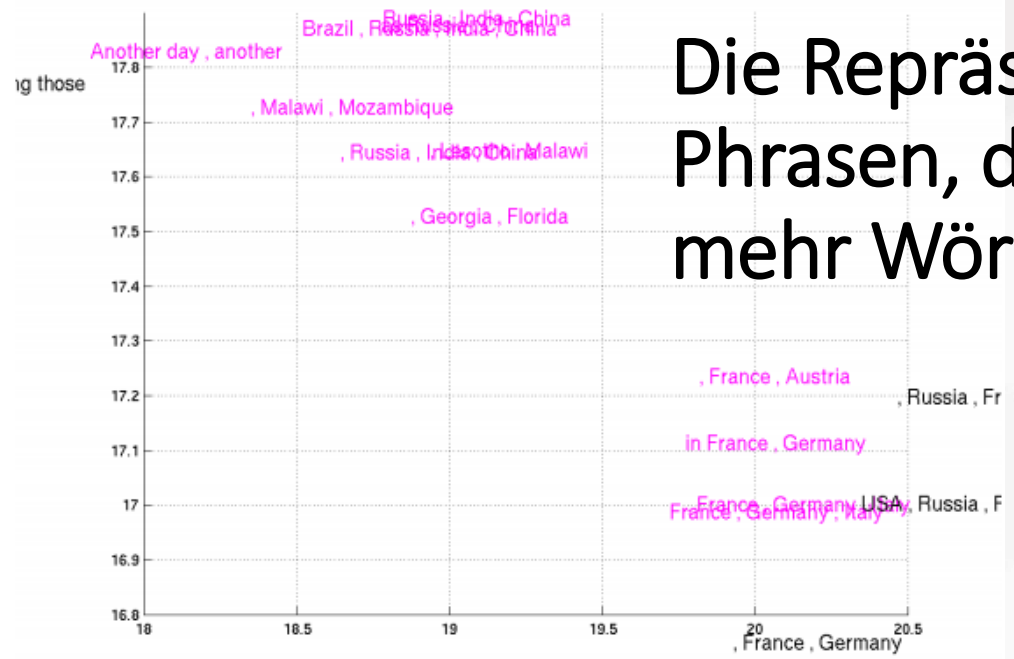

France
China
Russian
French
English
Germany
Ontario
Iraq
Japan
British
North
South


second
third
three
1
10
8
9
20
30
25
27
31
22
24
100
50
40
6
14
11
12
six
four
7
five


August
April
October
June
September
May
March
July
January
November
February
December

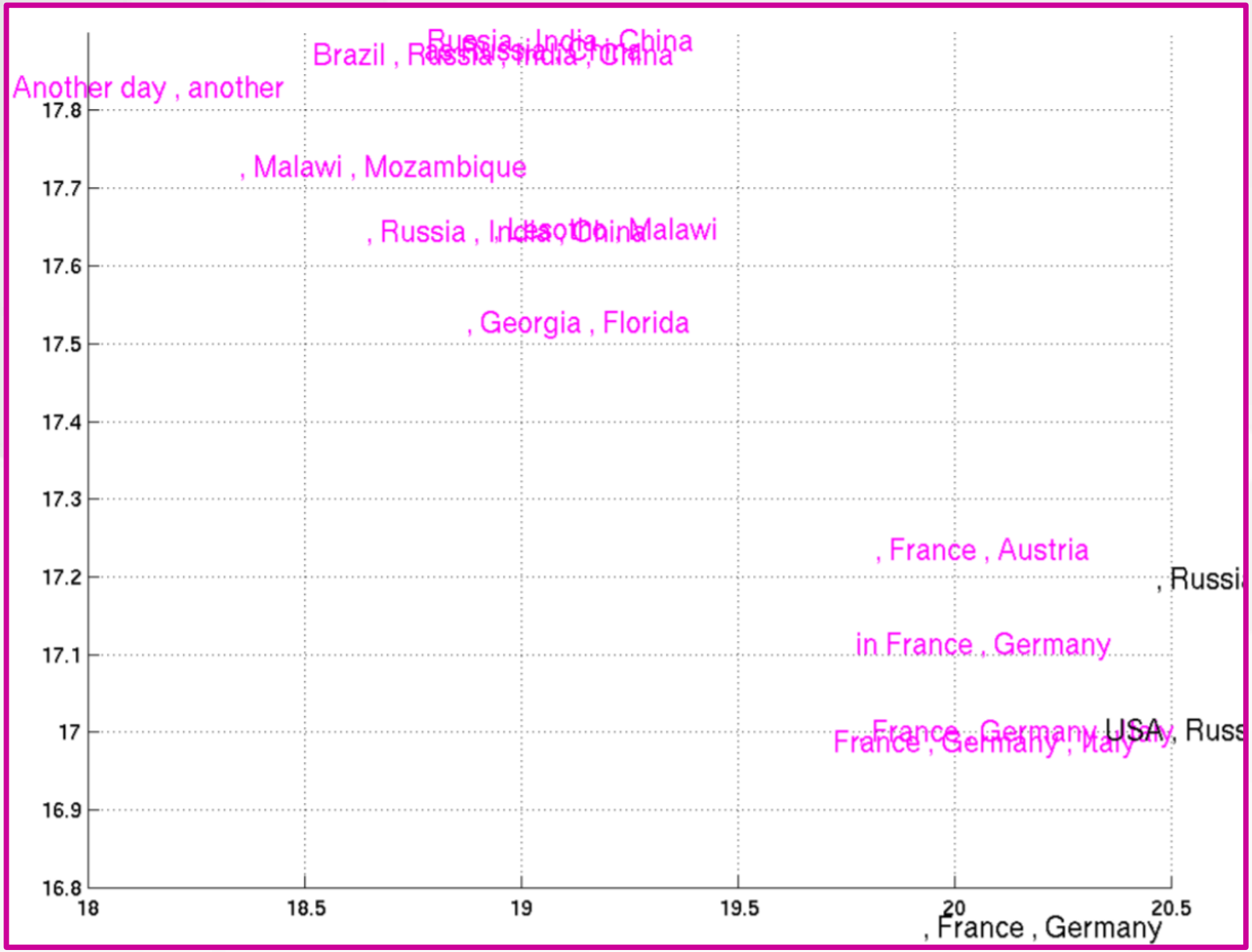




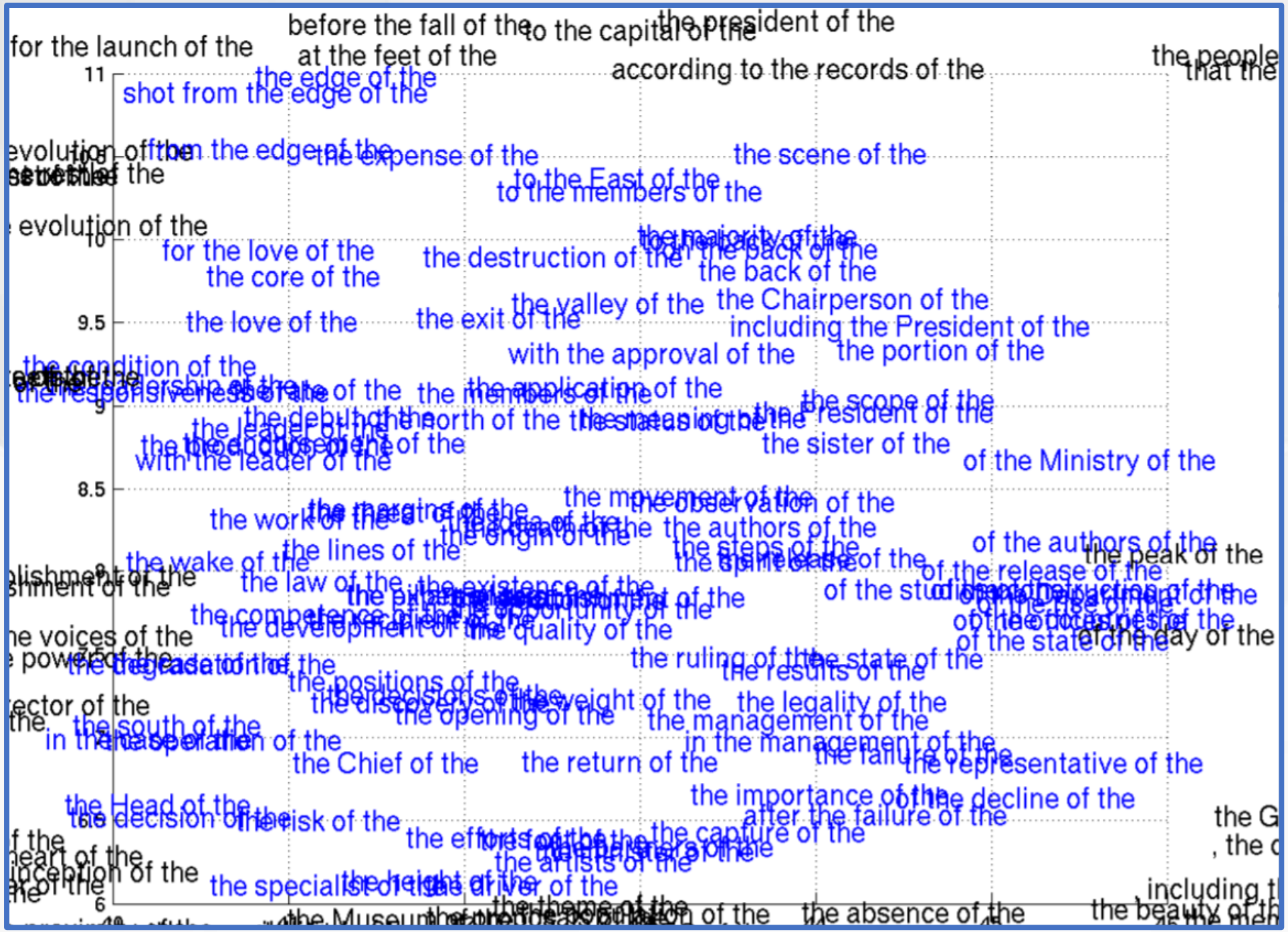


Die Repräsentationen der Phrasen, die aus vier oder mehr Wörtern besteht.

Syntaktisch ähnlich (über die Dauer der Zeit)



Semantisch ähnlich



Syntaktisch ähnlich

Agenda

- Introduction
- RNN Encoder-Decoder
  - Recurrent Neural Networks
  - RNN Encoder–Decoder
  - Hidden Unit that Adaptively Remembers and Forgets
- Statistical Machine Translation
  - Definition and examples of SMT
  - Scoring Phrase Pairs with RNN Encoder–Decoder
- Experiments
  - Data and Baseline System
  - Quantitative Analysis
  - Qualitative Analysis
  - Word and Phrase Representations
- Conclusion & Outlook

Zusammenfassung

- RNN Encoder–Decoder
  → Mapping von einer Sequenz beliebiger Länge zu einer anderen Sequenz.
  → Score & Generiere eine Zielsequenz.
- Hidden units
  → Reset-gates und Update-gates enthält.
- Das neue Modell
  → Gute Leistung und höhere BLEU-Score.

Ausblick

- großes Potenzial, ersetzen die ganze der Phrasentabelle
- zu anderen Anwendungen wie Sprachtranskription

Vielen Dank für Ihre Aufmerksamkeit!

Literatur

- https://arxiv.org/pdf/1406.1078.pdf
- https://en.wikipedia.org/wiki/Recurrent_neural_network
- https://en.wikipedia.org/wiki/Nonlinear_system
- https://en.wikipedia.org/wiki/Logistic_function
- https://en.wikipedia.org/wiki/Statistical_machine_translation
- https://en.wikipedia.org/wiki/Google_Translate
- http://www.statmt.org/wpt05/mt-shared-task/
- https://en.wikipedia.org/wiki/BLEU
- https://en.wikipedia.org/wiki/Neural_machine_translation
- http://statmt.org/wmt14/translation-task.html
- https://www.quora.com/What-is-the-meaning-of-low-rank-matrix