Neural Natural Language Processing in WebLicht

State-of-the-Art and Future Perspectives

Daniël de Kok

What is WebLicht?

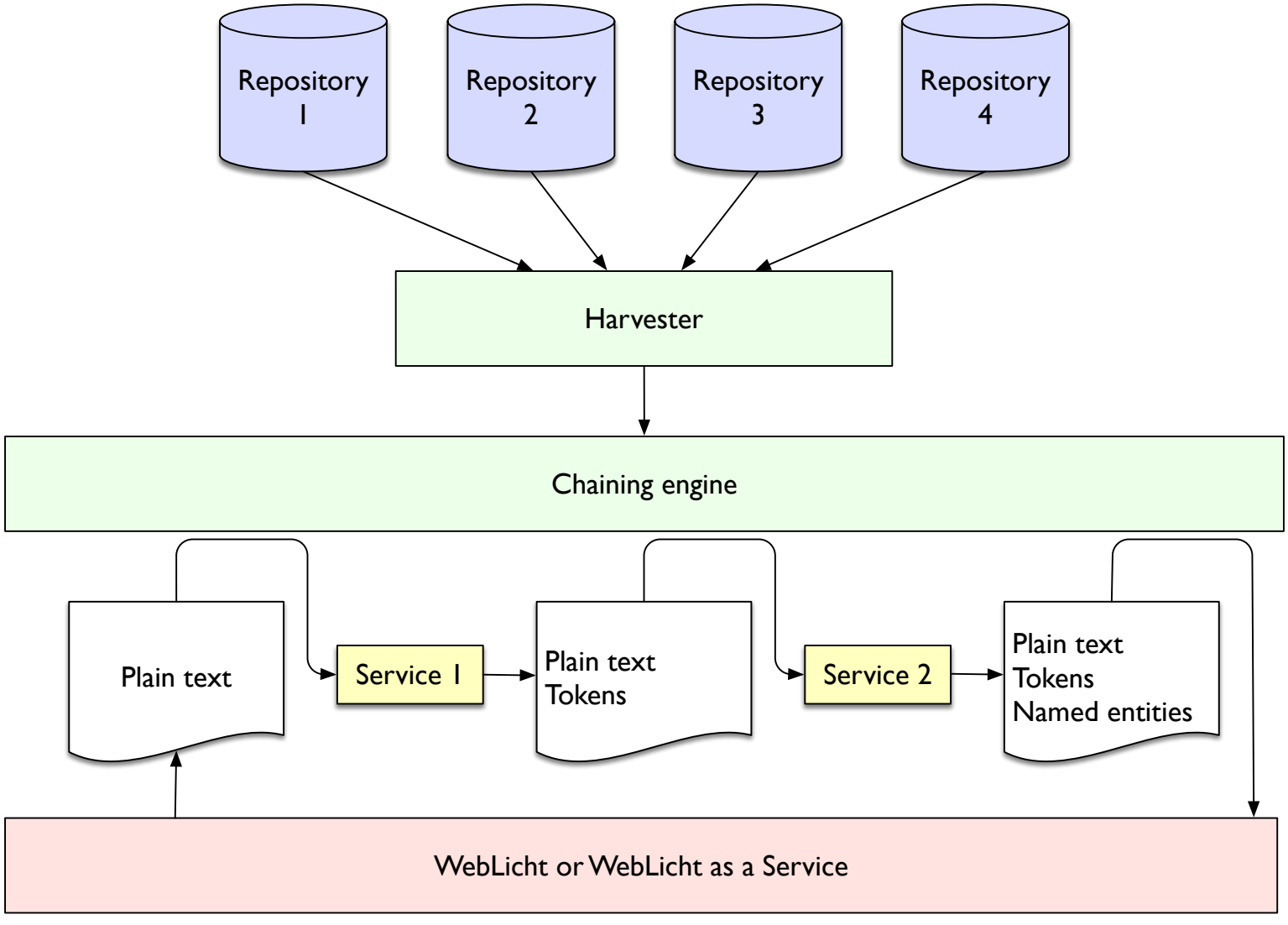
H.264

What is WebLicht?

H.264

Sherlock Holmes[PER] ist eine 1886 vom britischen Schriftsteller Arthur Conan Doyle[PER] geschaffene Kunstfigur .

WebLicht architecture



Text Corpus Format (TCF)

- XML data exchange format
- Annotations stored layer-wise
- Used by WebLicht webservices
- Tools for visualization, such as TüNDRA
- Tools for annotation, such as WebAnno

Text Corpus Format (TCF)

- XML data exchange format
- Annotations stored layer-wise
- Used by WebLicht webservices
- Tools for visualization, such as TüNDRA
- Tools for annotation, such as WebAnno
- Many supported annotation layers, including:
  - Tokenization
  - Lemmas
  - Constituency & Dependency Parsing
  - Morphology
  - Named Entitities
  - References (anaphoric, cataphoric, coreference)
  - Geographical locations
  - Text structure

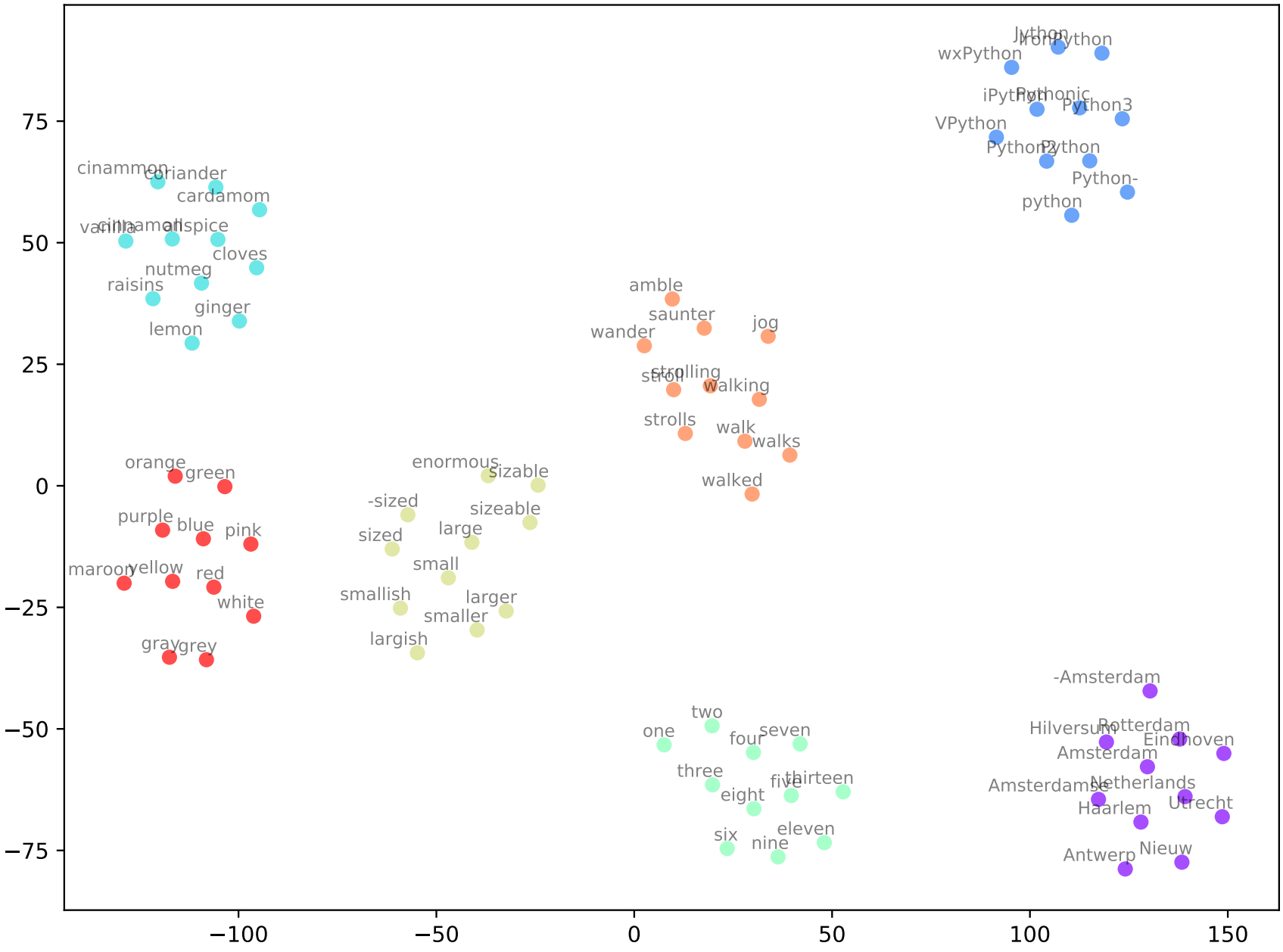
State-of-the-Art

State-of-the-Art: Neural language processing

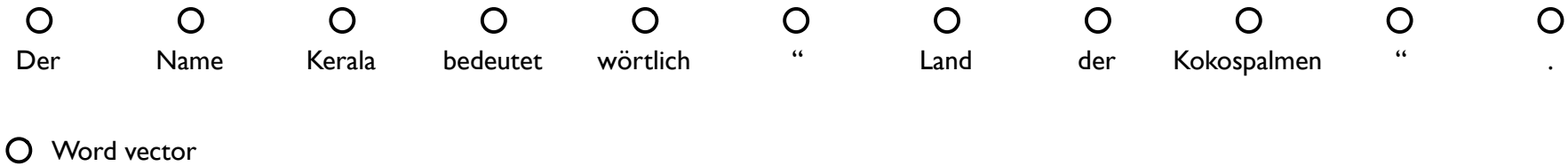
Key ingredients:

- Vector representations of words
- Deep neural networks

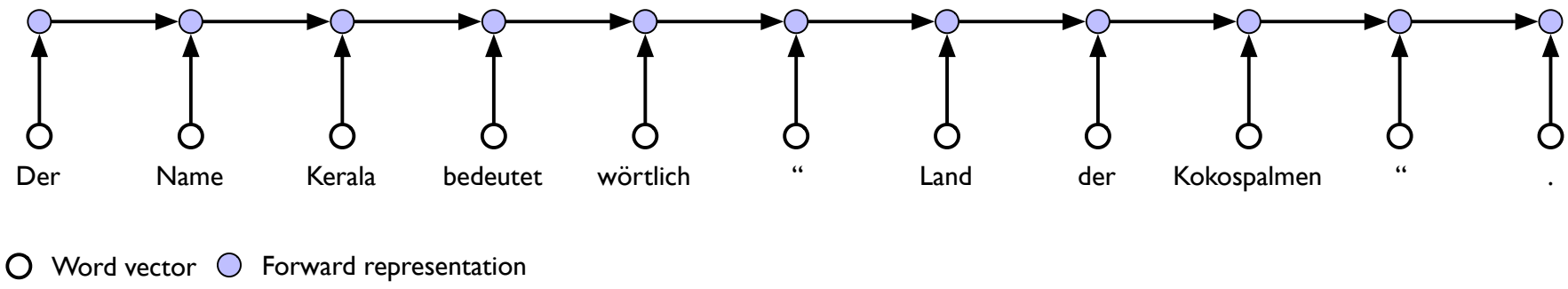
Words as vectors



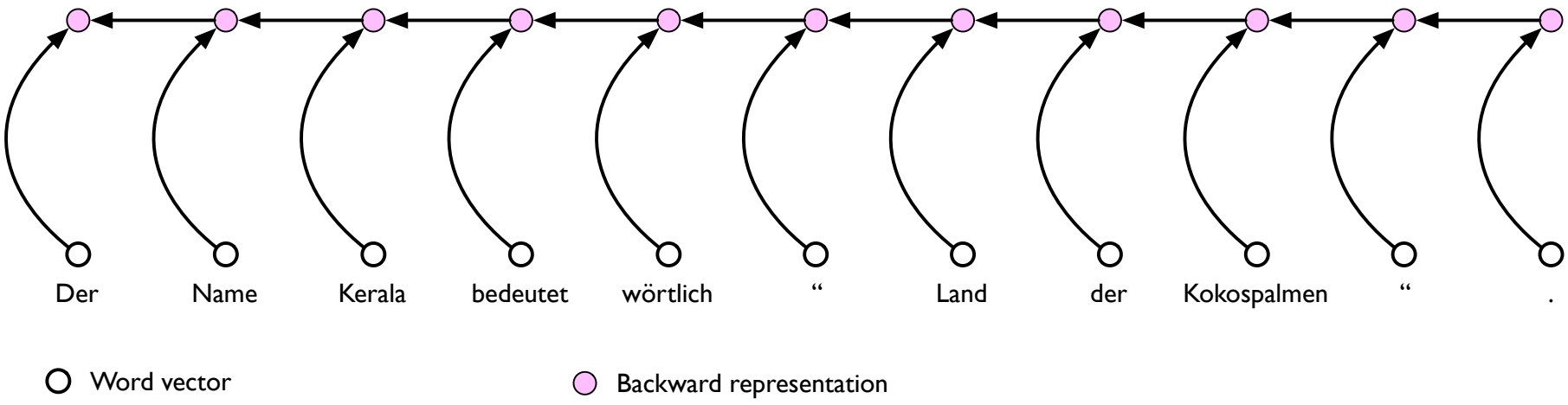
Recurrent neural networks (RNN)



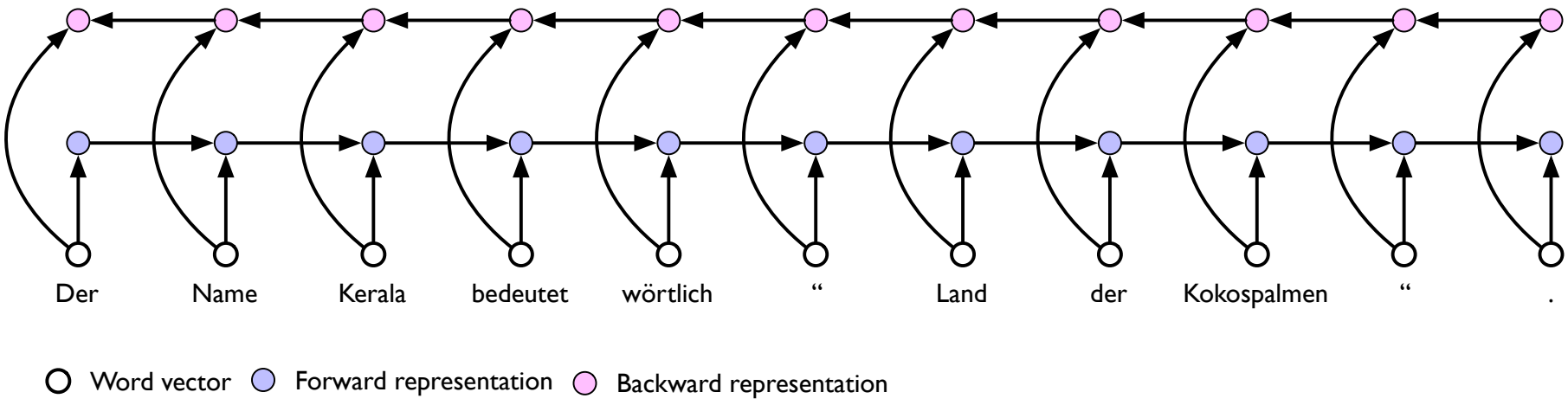
Recurrent neural networks (RNN)



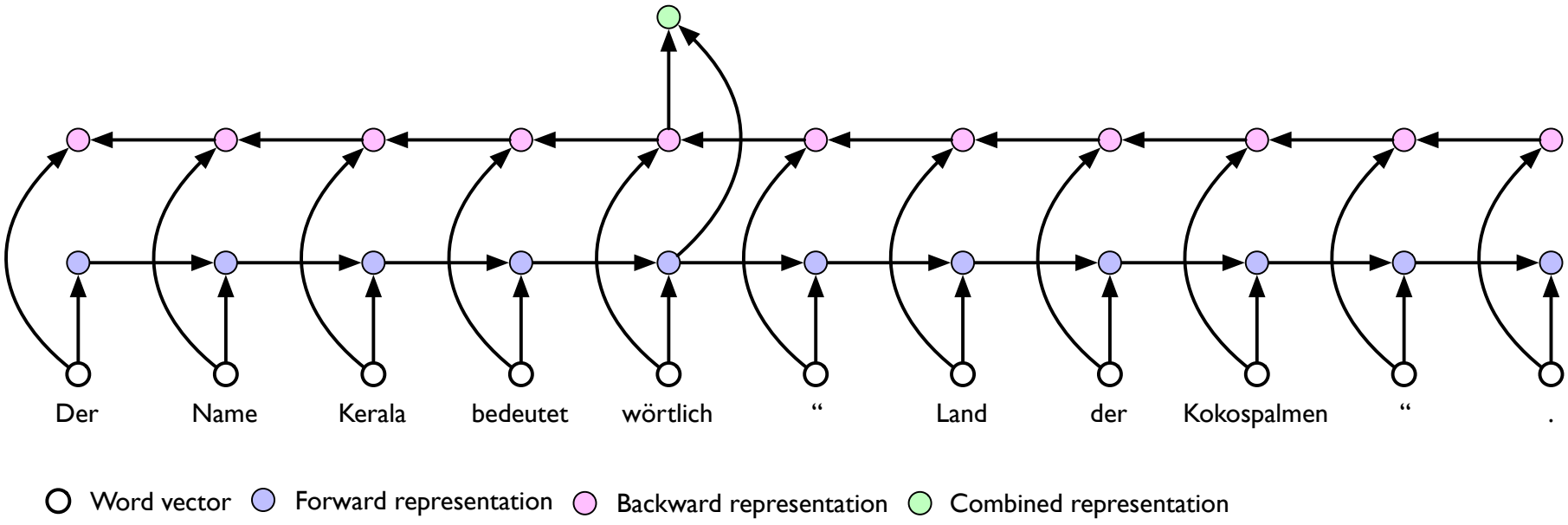
Recurrent neural networks (RNN)



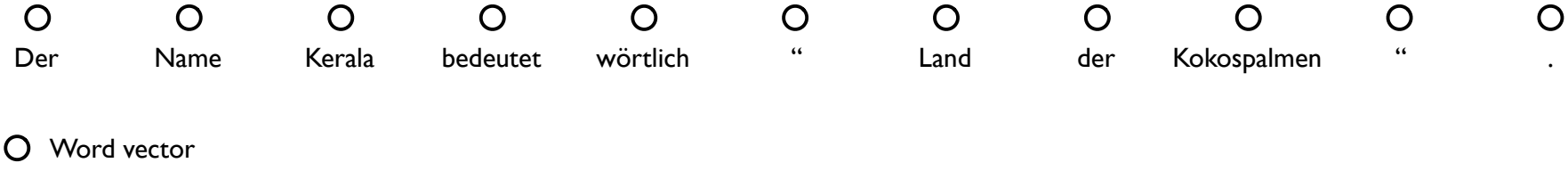
Recurrent neural networks (RNN)



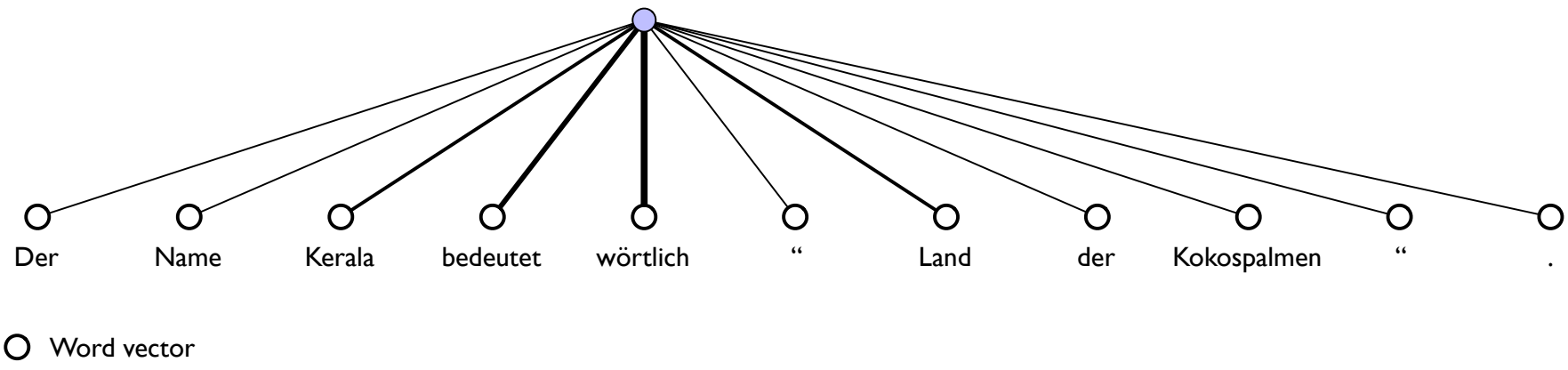
Recurrent neural networks (RNN)



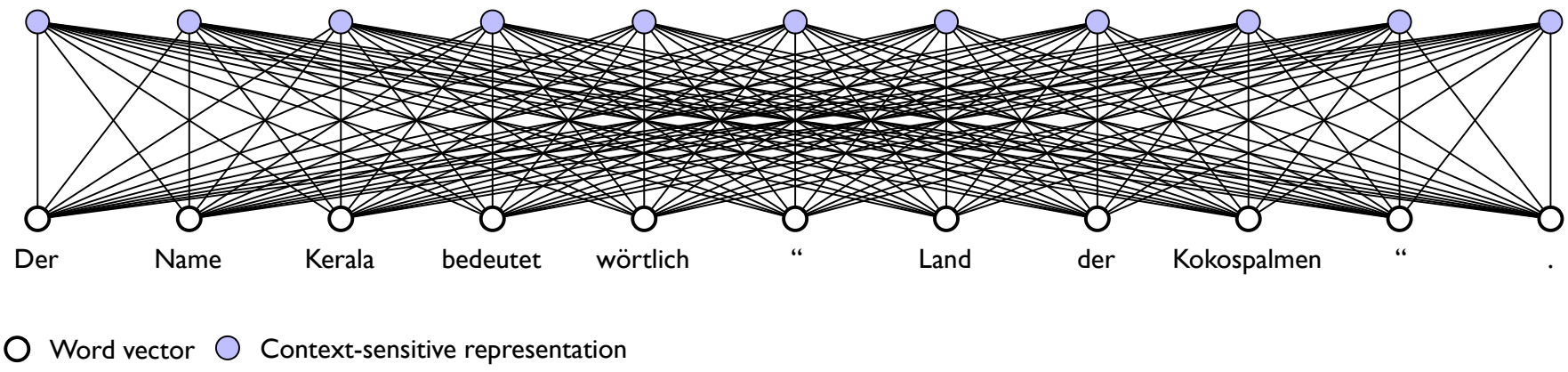
Transformer



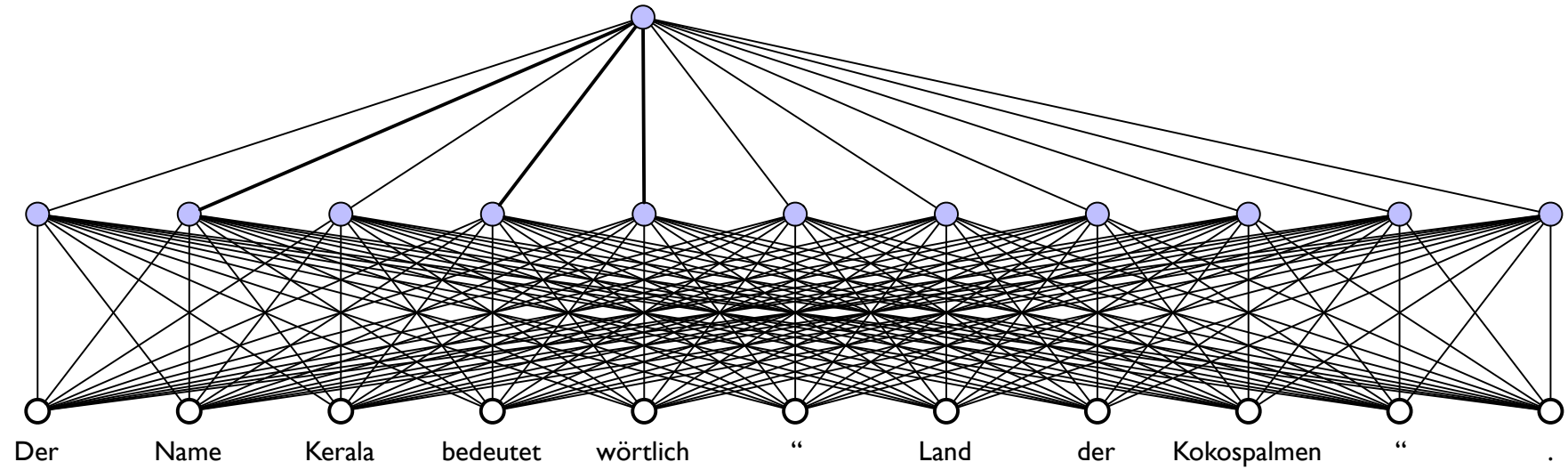
Transformer



Transformer



Transformer



Improvements in syntactic tasks

[1] Joint work with Erik Schill.
[2] Joint work with Tobias Pütz.

Improvements in syntactic tasks

Part-of-speech tagging[1]

VVFIN	**ART**	**NN**	**PTKVZ**	**$.**
Wacht	die	Opposition	auf	?

[1] Joint work with Erik Schill.

[2] Joint work with Tobias Pütz.

Improvements in syntactic tasks

Part-of-speech tagging[1]

VVFIN	**ART**	**NN**	**PTKVZ**	**$.**
Wacht	die	Opposition	auf	?

Parser	**Accuracy**
HMM	97.29
RNN	99.03

[1] Joint work with Erik Schill.

[2] Joint work with Tobias Pütz.

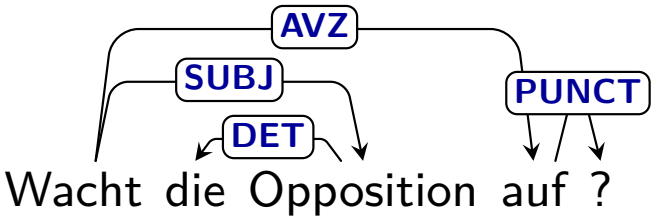
Improvements in syntactic tasks

Part-of-speech tagging[1]

VVFIN	**ART**	**NN**	**PTKVZ**	**$.**
Wacht	die	Opposition	auf	?

Parser	Accuracy
HMM	97.29
RNN	99.03

Dependency parsing[2]



[1] Joint work with Erik Schill.

[2] Joint work with Tobias Pütz.

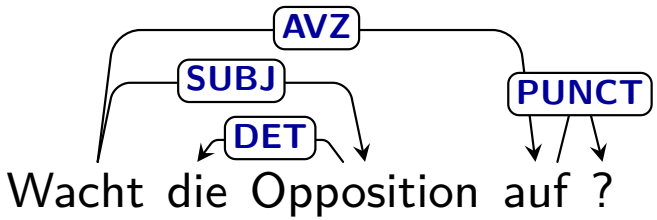
Improvements in syntactic tasks

Part-of-speech tagging[1]

VVFIN	ART	NN	PTKVZ	$.
Wacht	die	Opposition	auf	?

Parser	Accuracy
HMM	97.29
RNN	99.03

Dependency parsing[2]



Tagger	Accuracy
Linear SVM	89.9
Feed forward NN	91.8
RNN	94.4
Transformer	94.7
RNN + pretraining	96.1
Transformer + pretraining	96.3

[1] Joint work with Erik Schill.

[2] Joint work with Tobias Pütz.

Conclusion

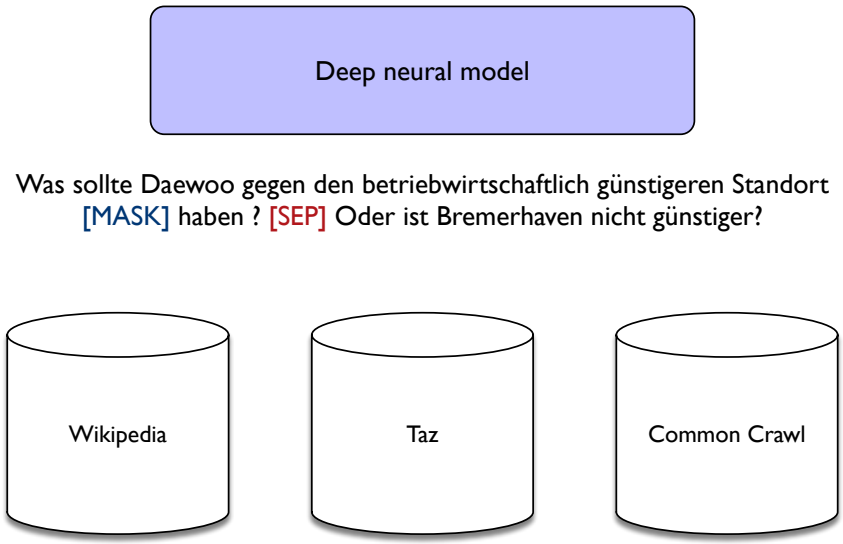
- WebLicht allows you to build text analysis piplines: https://weblicht.sfs.uni-tuebingen.de
- WebLicht can be integrated in other applications using WebLicht as a Service: https://weblicht.sfs.uni-tuebingen.de/WaaS/
- State-of-the-Art: annotations tools using neural networks

Thank you!

Future perspectives

Pretraining (Peters et al., 2018 & Devlin et al., 2019)

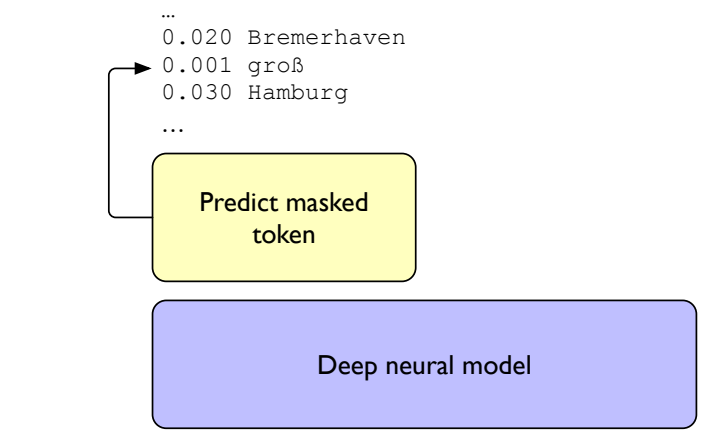
Pretraining



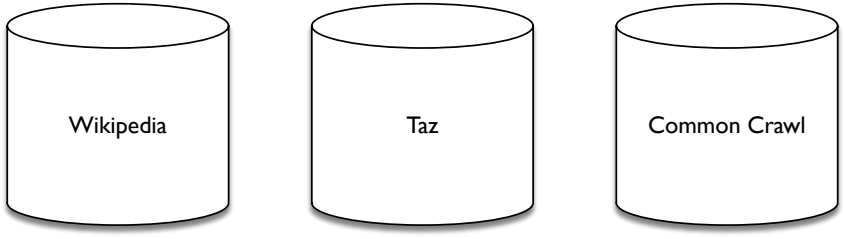
Finetuning

Pretraining (Peters et al., 2018 & Devlin et al., 2019)

Pretraining



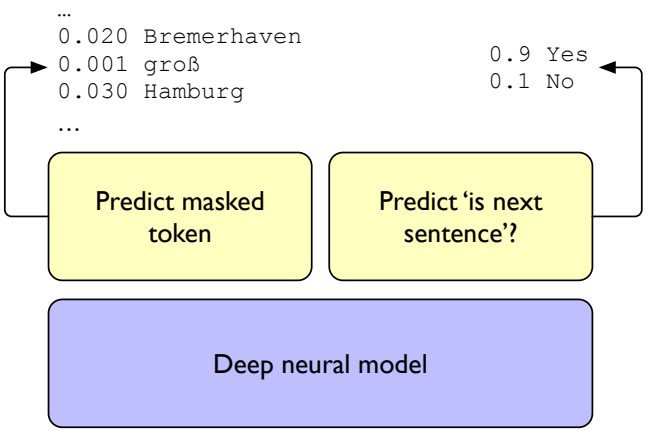
Was sollte Daewoo gegen den betriebwirtschaftlich günstigeren Standort [MASK] haben ? [SEP] Oder ist Bremerhaven nicht günstiger?



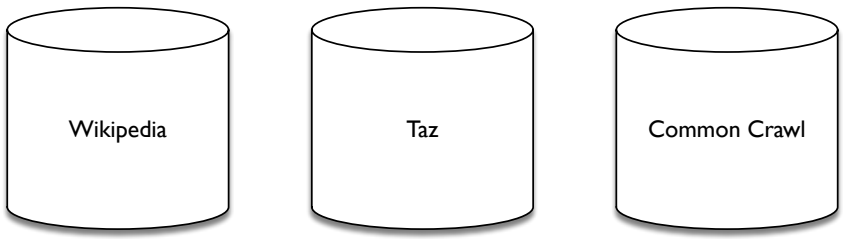
Finetuning

Pretraining (Peters et al., 2018 & Devlin et al., 2019)

Pretraining



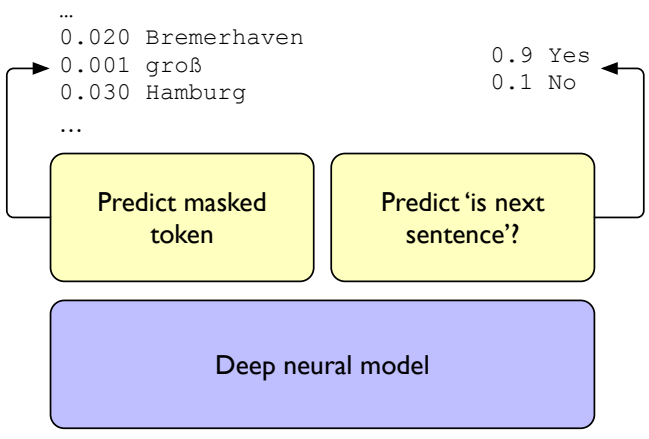
Was sollte Daewoo gegen den betriebwirtschaftlich günstigeren Standort [MASK] haben ? [SEP] Oder ist Bremerhaven nicht günstiger?



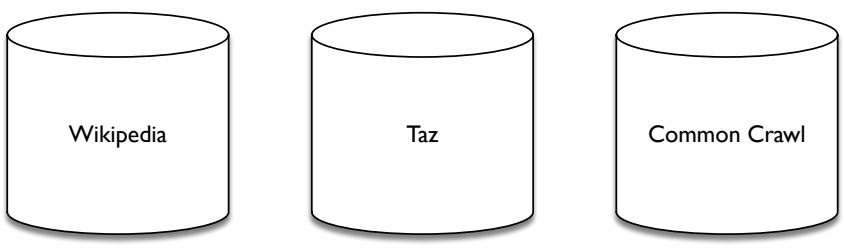
Finetuning

Pretraining (Peters et al., 2018 & Devlin et al., 2019)

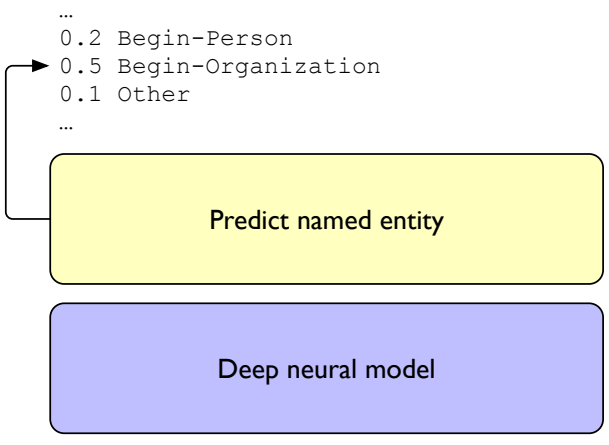
Pretraining



Was sollte Daewoo gegen den betriebwirtschaftlich günstigeren Standort [MASK] haben ? [SEP] Oder ist Bremerhaven nicht günstiger?



Finetuning



Was sollte Daewoo gegen den betriebwirtschaftlich günstigeren Standort Bremerhaven haben ?

Pretraining holds promise for text analysis

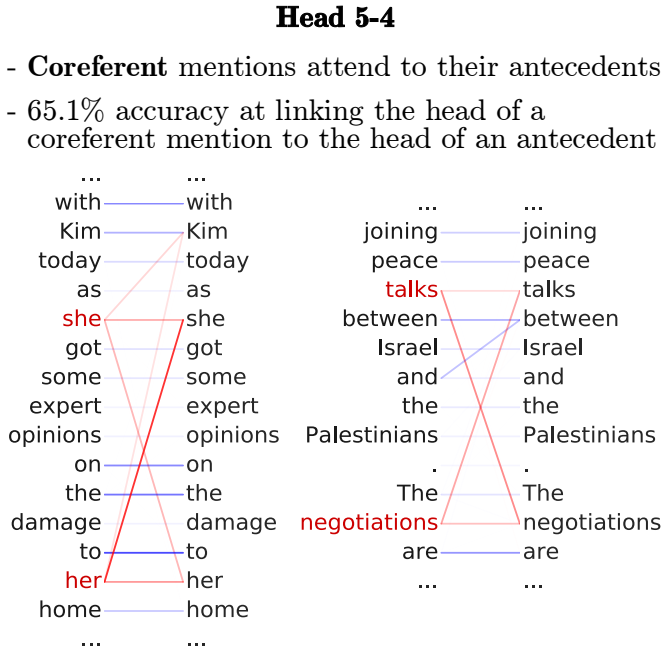
- Pretraining models seem to pick up a certain amount of syntax (Tenney et al., 2019 & Clark et al., 2019)

Pretraining holds promise for text analysis

- Pretraining models seem to pick up a certain amount of syntax (Tenney et al., 2019 & Clark et al., 2019)
- Pretraining models with fine tuning shown to work well on textual entailment, paraphrase detection, and question-answer pair classification. (Devlin et al., 2019)

Pretraining holds promise for text analysis

- Pretraining models seem to pick up a certain amount of syntax (Tenney et al., 2019 & Clark et al., 2019)
- Pretraining models with fine tuning shown to work well on textual entailment, paraphrase detection, and question-answer pair classification. (Devlin et al., 2019)



(Source: Clark et al., 2019)