Grundlagen Kontextualisierter Wordembeddings

Erik Fäßler

Jena University Language & Information Engineering (JULIE)
Lab Friedrich-Schiller-University Jena, Germany

http://www.julielab.de

Language Models

Forward-Model

given the history $(t_1, ..., t_{k-1})$:

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

Backward-Model

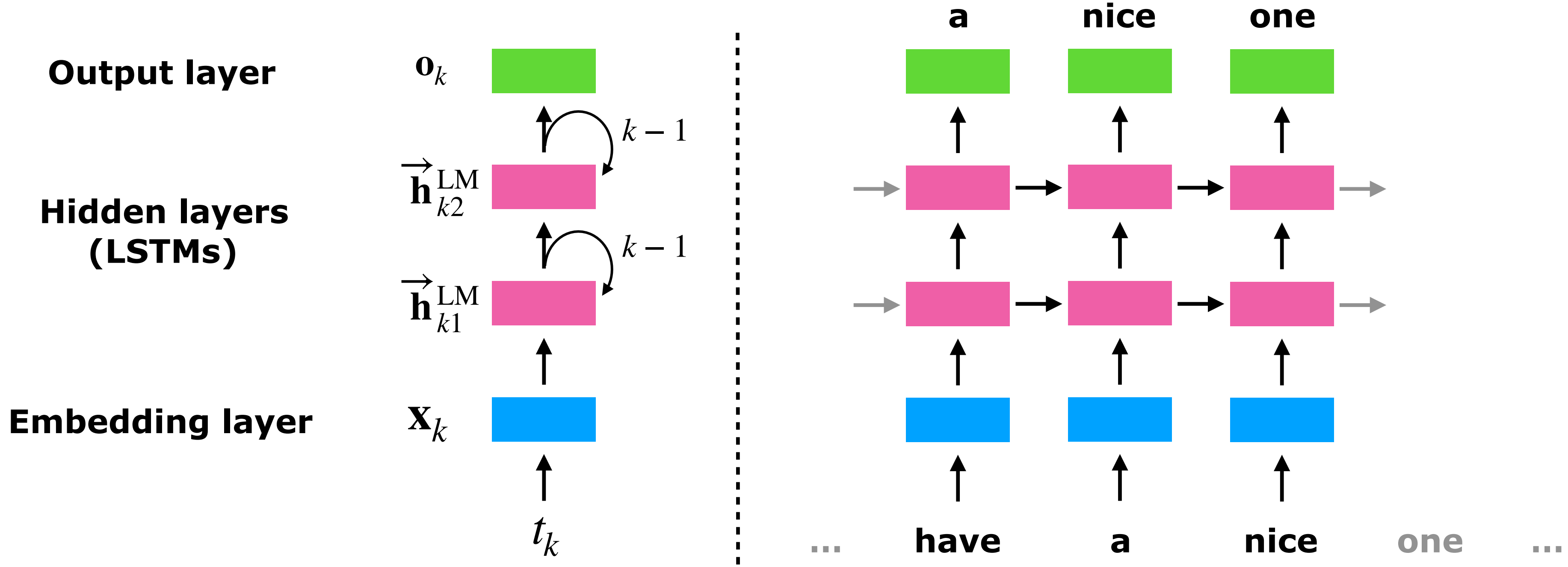
$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

Deep LM

With long short term memory (LSTM) network, predicting the next words in both directions to build biLMs

The forward LM architecture

Expanded in the forward direction of *k*

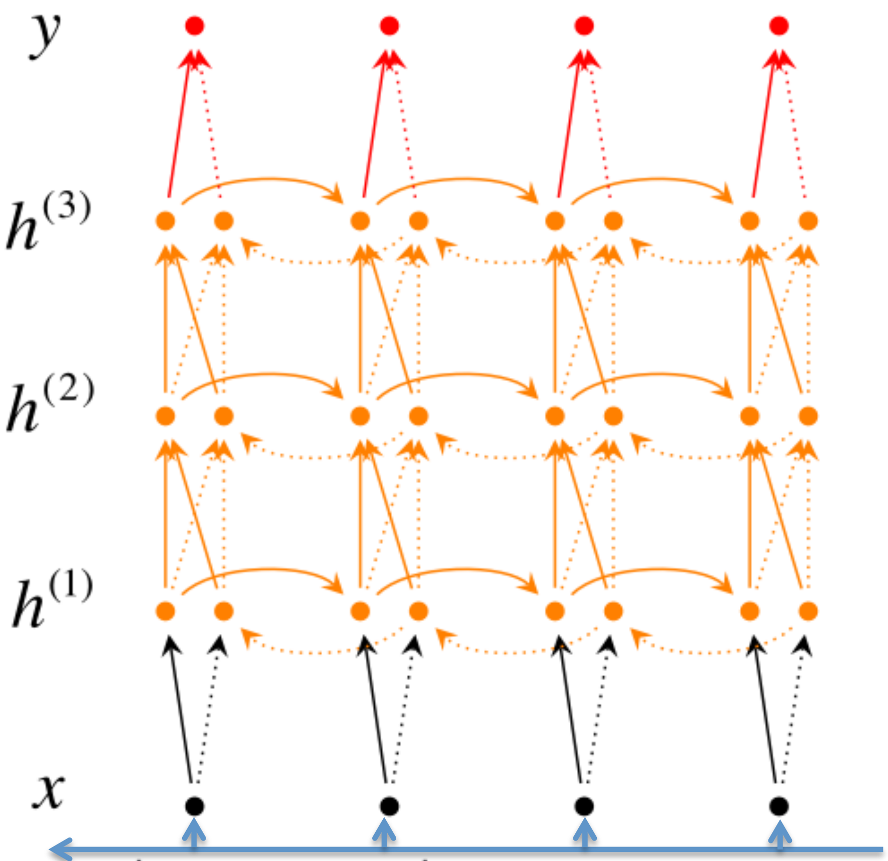


https://de.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018

Deep Bidirectional LM

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/syllabus.html, Lecture 8 (Feb 2), Slide 46

Deep Bidirectional RNNs



$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

$$y_t = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

static word embeddings (word2vec, glove, fasttext, ...)

Each memory layer passes an intermediate sequential representation to the next.

ELMo (Peters et al.)

- BI-LSTM-LM mit L=2 Layern
- Für jeden Token-position k extrahiere
  - den statischen Embeddingvector (CNN character n-grams!)
  - die forward und backword biLM Zwischenrepräsentationen

$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

0 = Token layer = x_k^{LM}

- Linearkombination der Vektoren
- Task-spezifische Parameter

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.$$

- Ersetze im eigentlichen Aufgaben-KNN die statischen Embeddings durch ELMo

ELMo Vectors

Jena University Language & Information Engineering Lab

Structure

Each token t_k

L-layer biLM computes 2L+1 representations

$\overrightarrow{h}^{LM}_{k,j}$

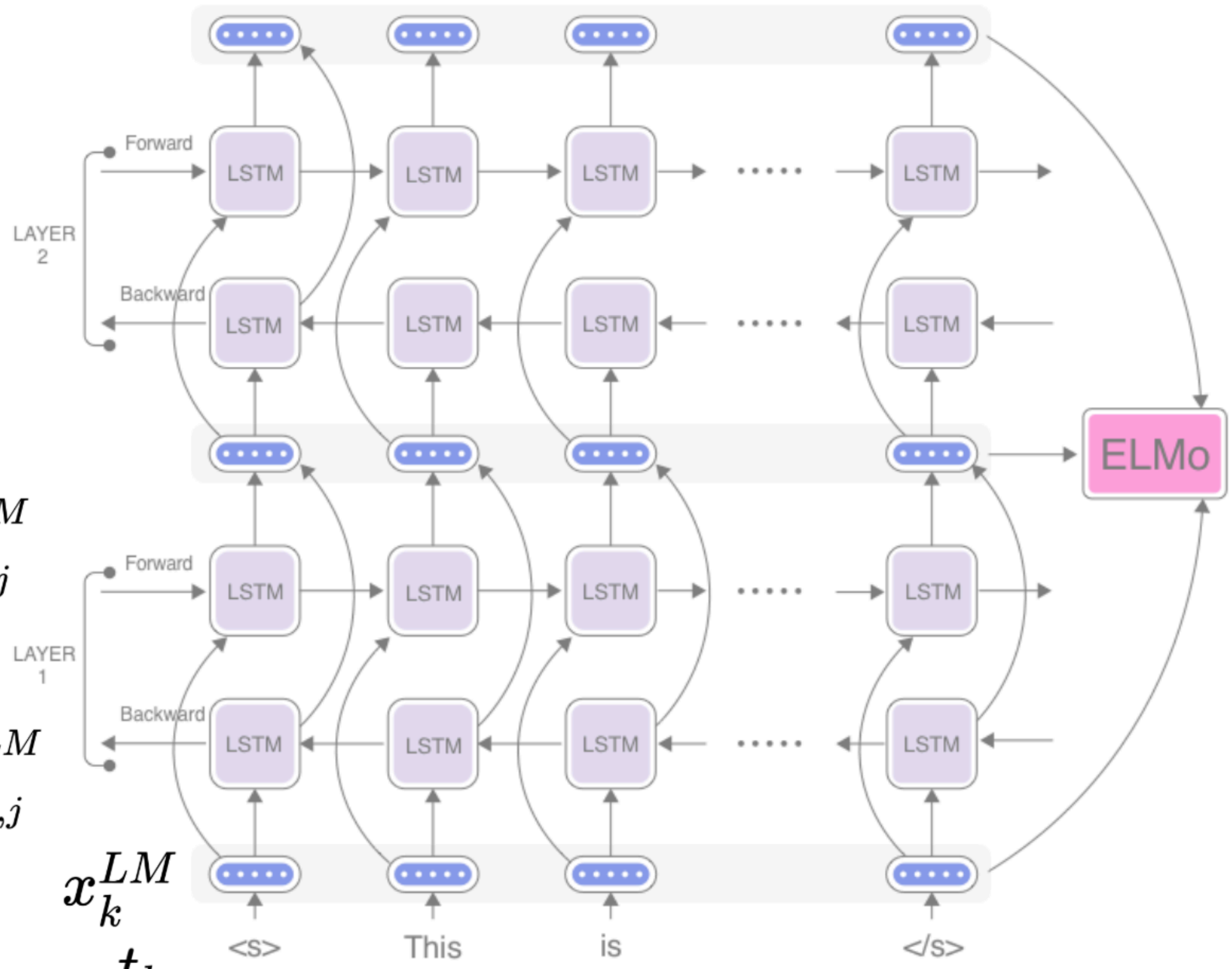
k is the k-th token

$\overleftarrow{h}^{LM}_{k,j}$

j is the j-th biLM layer

x^{LM}_{k}

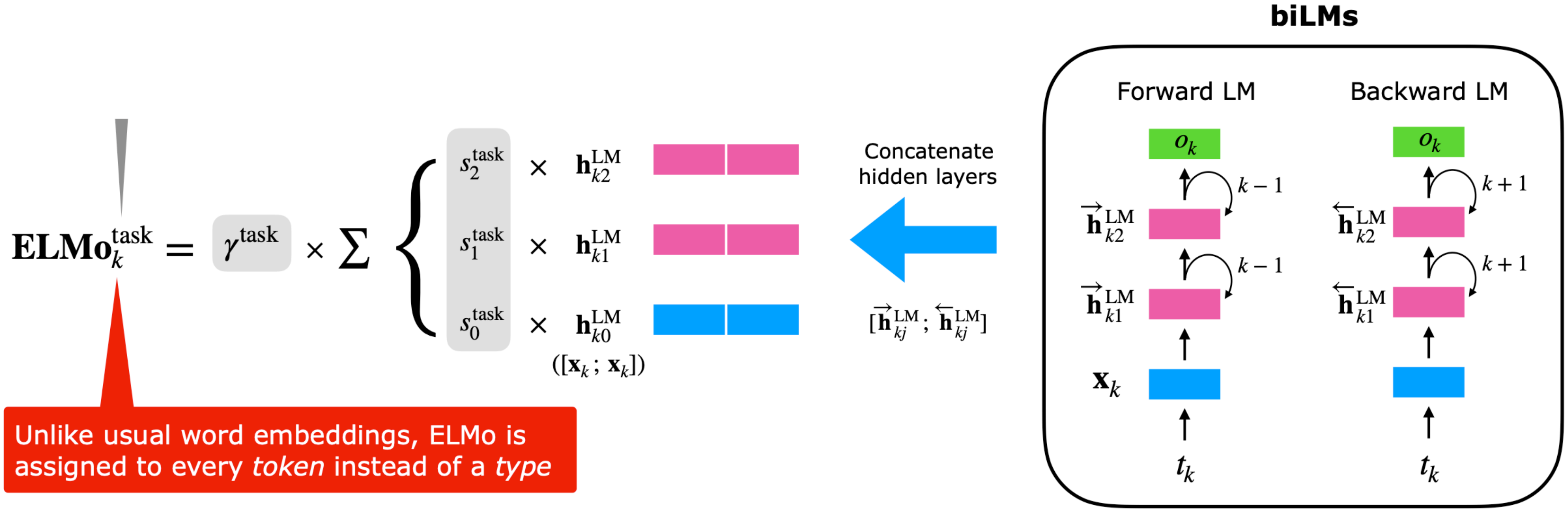
t_k



https://ireneli.eu/2018/12/17/elmo-in-practice/

ELMo Model

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)



37

https://de.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018

- Question answering (Stanford Question Answering Dataset, SQuAD)
- Textual entailment (Stanford Natural Language Inference (SNLI) corpus
- Semantic role labeling (OntoNotes)
- Coreference solution (OntoNotes)
- Named Entity Extraction (CoNLL 2003 NER)
- Sentiment analysis

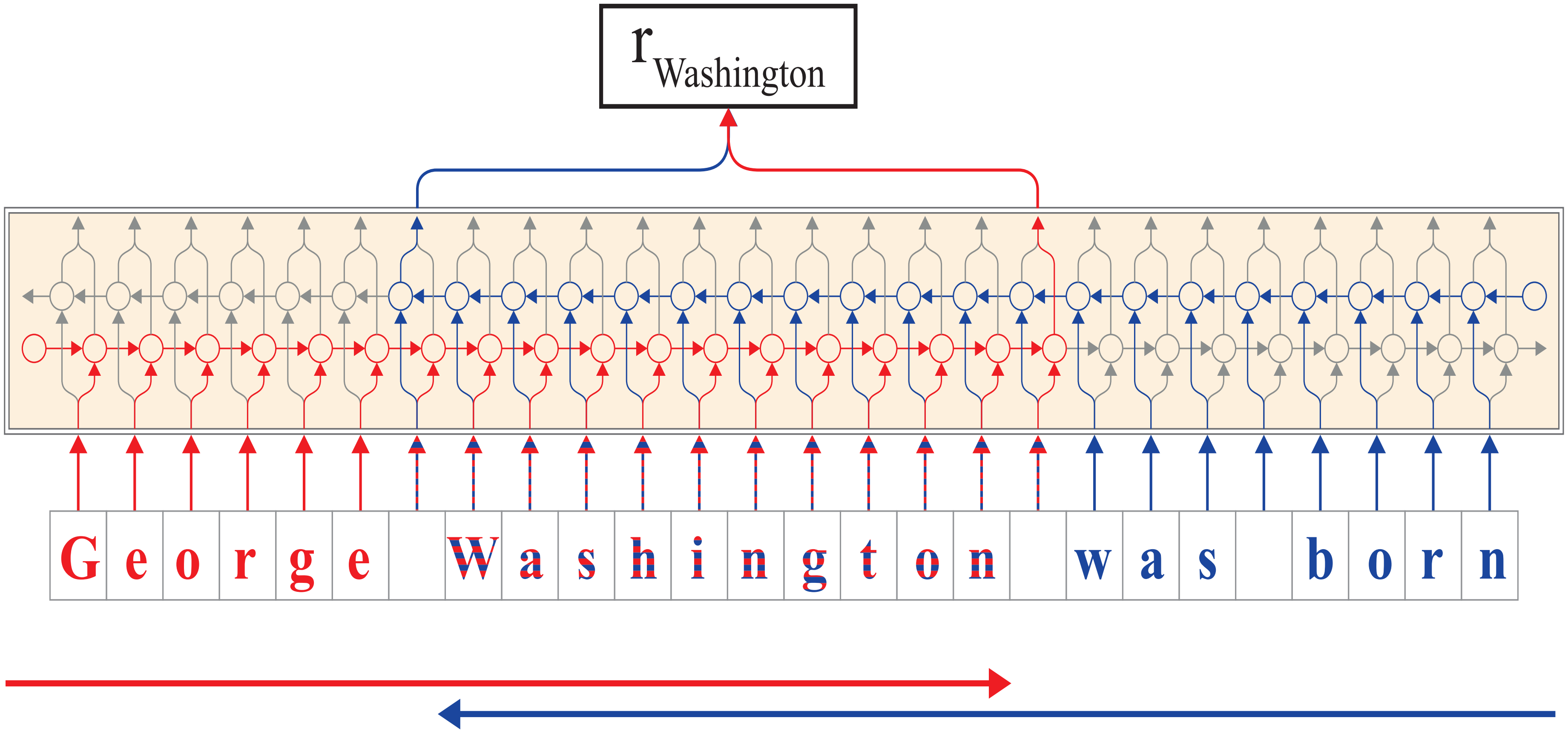
ELMo Evaluation

nicht-triviale KNN Architekturen

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Flair (Akbik et al.)

- Basiert ebenfalls auf biLMs
- Allerdings nicht auf Token- sondern auf Zeichenebene
- Bislang nur Sequenzentagging
- Im Flair Framework gibt es auch Dokumentenklassifikation
- Flair embedding für token k:
  - Forward: LM Hidden State nach letztem Token Zeichen
  - Backward: LM Hidden State vor erstem Zeichen
- Im Flair Paper und –Framework werden die unterschiedlichen Embeddings concateniert (gestackt)
  - Flair Forward + Backward + (GloVe | word2vec | fasttext | ...)


$r_{Washington}$
G e o r g e W a s h i n g t o n w a s b o r n

ELMo

Task	Language	Dataset	Flair	Previous best
Named Entity Recognition	English	Conll-03	**93.18** (F1)	*92.22 (Peters et al., 2018)*
Named Entity Recognition	English	Ontonotes	**89.3** (F1)	*86.28 (Chiu et al., 2016)*
Emerging Entity Detection	English	WNUT-17	**49.49** (F1)	*45.55 (Aguilar et al., 2018)*
Part-of-Speech tagging	English	WSJ	**97.85**	*97.64 (Choi, 2016)*
Chunking	English	Conll-2000	**96.72** (F1)	*96.36 (Peters et al., 2017)*
Named Entity Recognition	German	Conll-03	**88.27** (F1)	*78.76 (Lample et al., 2016)*
Named Entity Recognition	German	Germeval	**84.65** (F1)	*79.08 (Hänig et al, 2014)*
Named Entity Recognition	Dutch	Conll-03	**90.44** (F1)	*81.74 (Lample et al., 2016)*
Named Entity Recognition	Polish	PolEval-2018	**86.6** (F1) (Borchmann et al., 2018)	*85.1 (PolDeepNer)*