Automatic Text Simplification with CNNs

Dominik Pfütze, Henny Sluyter-Gäthje,
Simon Untergasser, Tim Patzelt

Outline

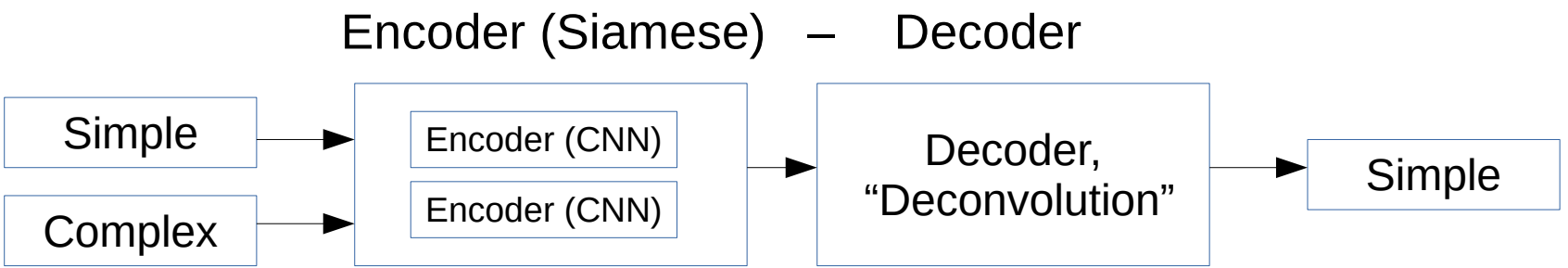
- Recap of our idea
- Methods and Results
  - Word embeddings
  - Encoder
  - Decoder
  - End-to-end model
- Conclusion

Recap of our approach

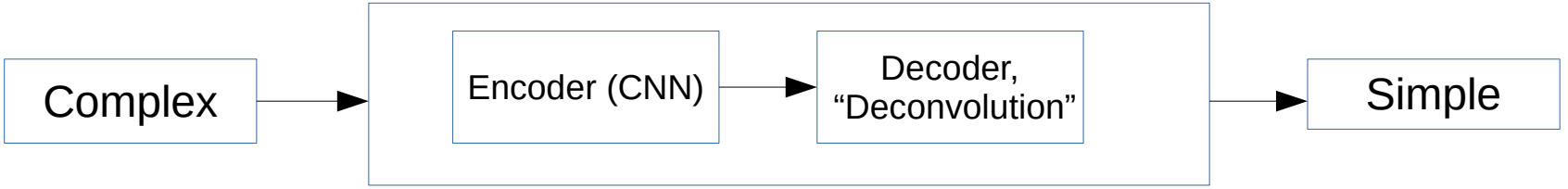
- Corpus: English and Simple English Wikipedia (EW-SEW data set)
  - 284.678 automatically aligned sentences[1]
  - Encoder: additionally same number of random sentence pairs from corpus
  - max. length of sentence = 50

[1] Hwang et al, 2015, Aligning sentences from standard wikipedia to simple wikipedia

Recap of our approach



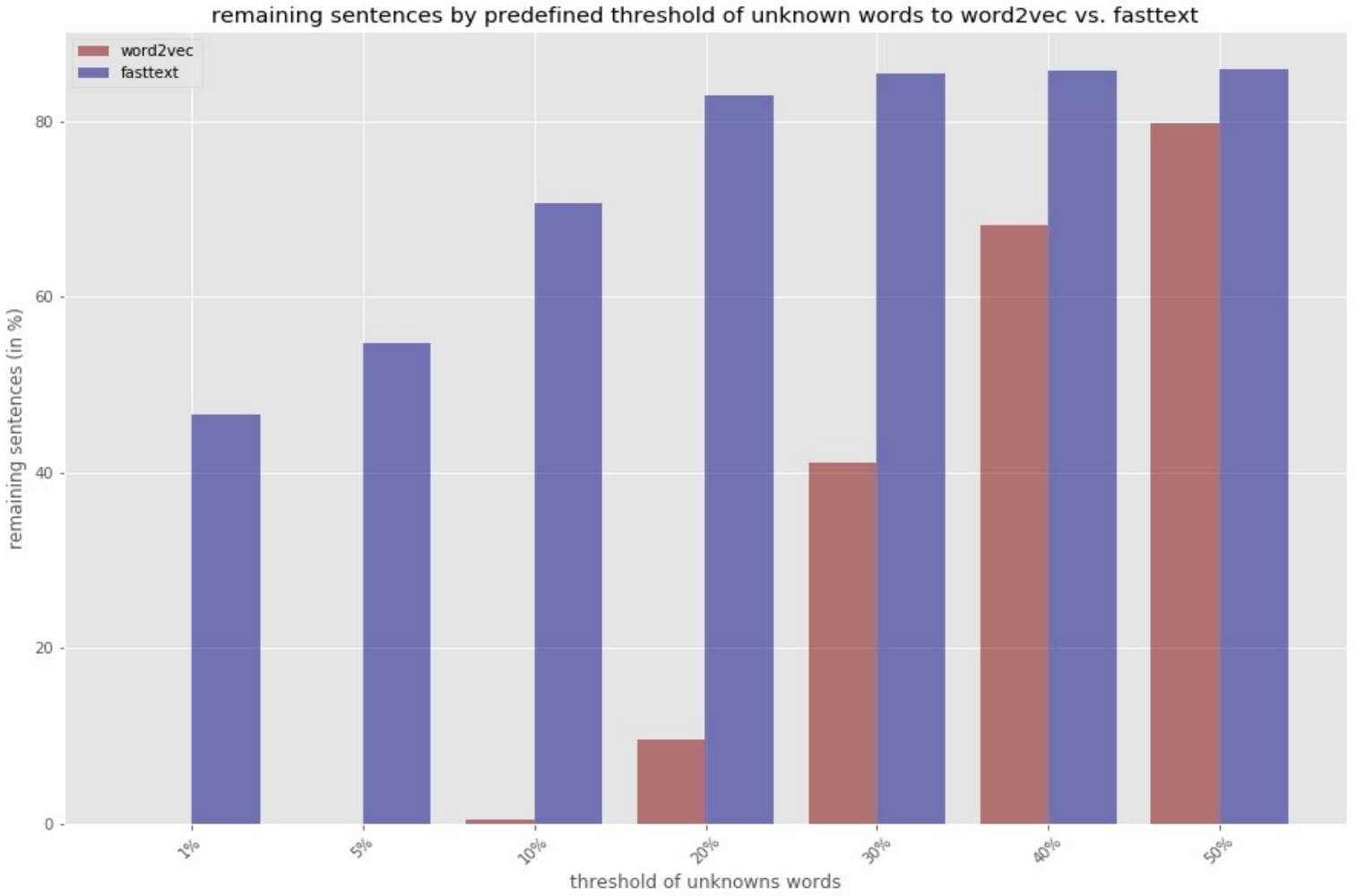
vs. End-to-End



Word embeddings

- Word2Vec[1] vs. FastText[2]
- With Word2Vec we had a lot of unknowns

[1] Mikolov et al., 2013, Efficient estimation of word representations in vector space
[2] Bojanewski et al., 2016, Enriching word vectors with subword information


remaining sentences by predefined threshold of unknown words to word2vec vs. fasttext
word2vec
fasttext
remaining sentences (in %)
0
20
40
60
80
1%
5%
10%
20%
30%
40%
50%
threshold of unknowns words

Word embeddings

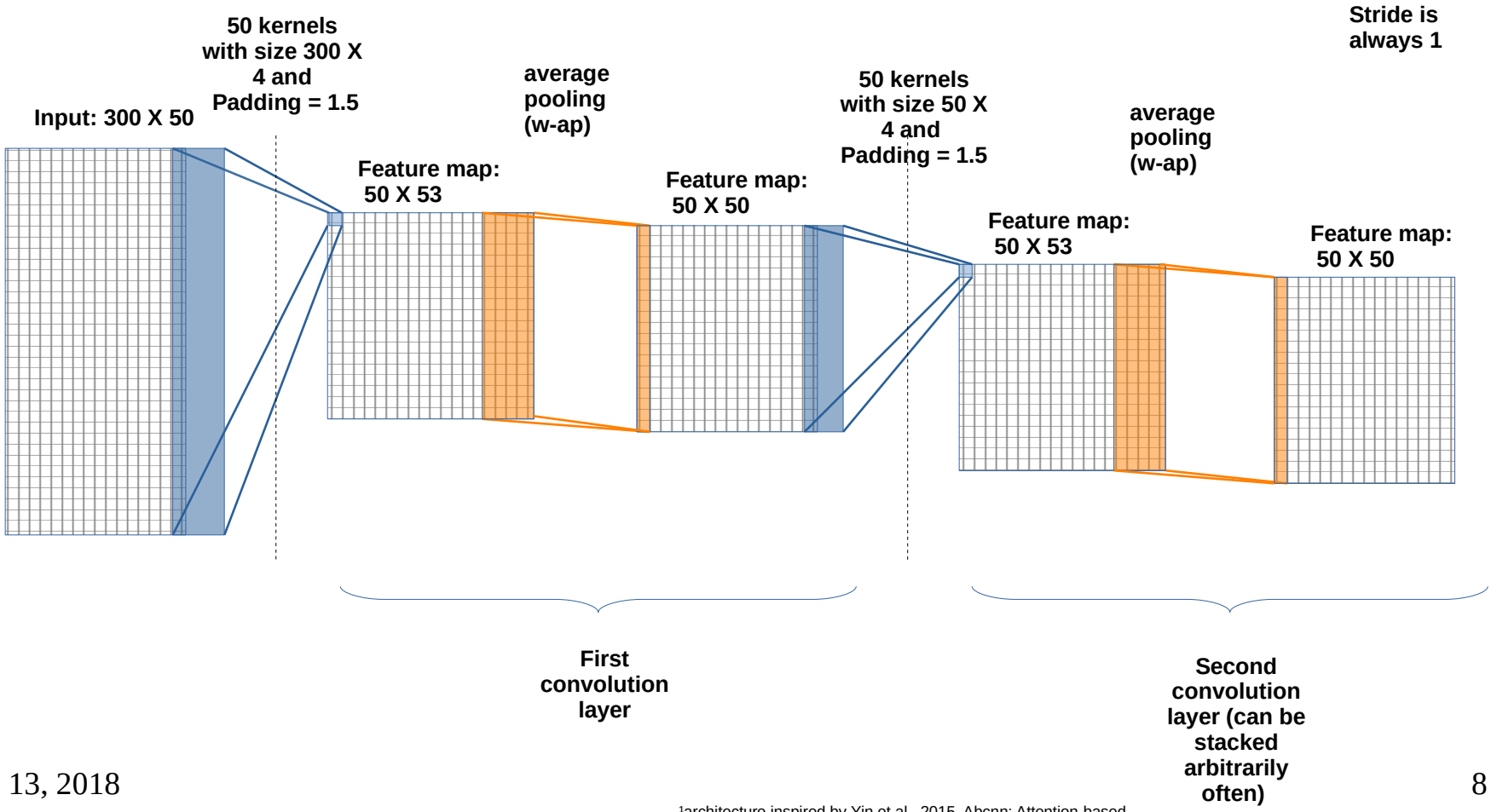
- Word2Vec[1] vs. FastText[2]

- With Word2Vec we had a lot of unknowns

  → switched to FastText working with character n-grams (3≤n≤6) instead of words

[1] Mikolov et al., 2013, Efficient estimation of word representations in vector space
[2] Bojanewski et al., 2016, Enriching word vectors with subword information

Encoder[1]



[1]architecture inspired by Yin et al., 2015, Abcnn: Attention-based convolutional neural network for modeling sentence pairs

Encoder

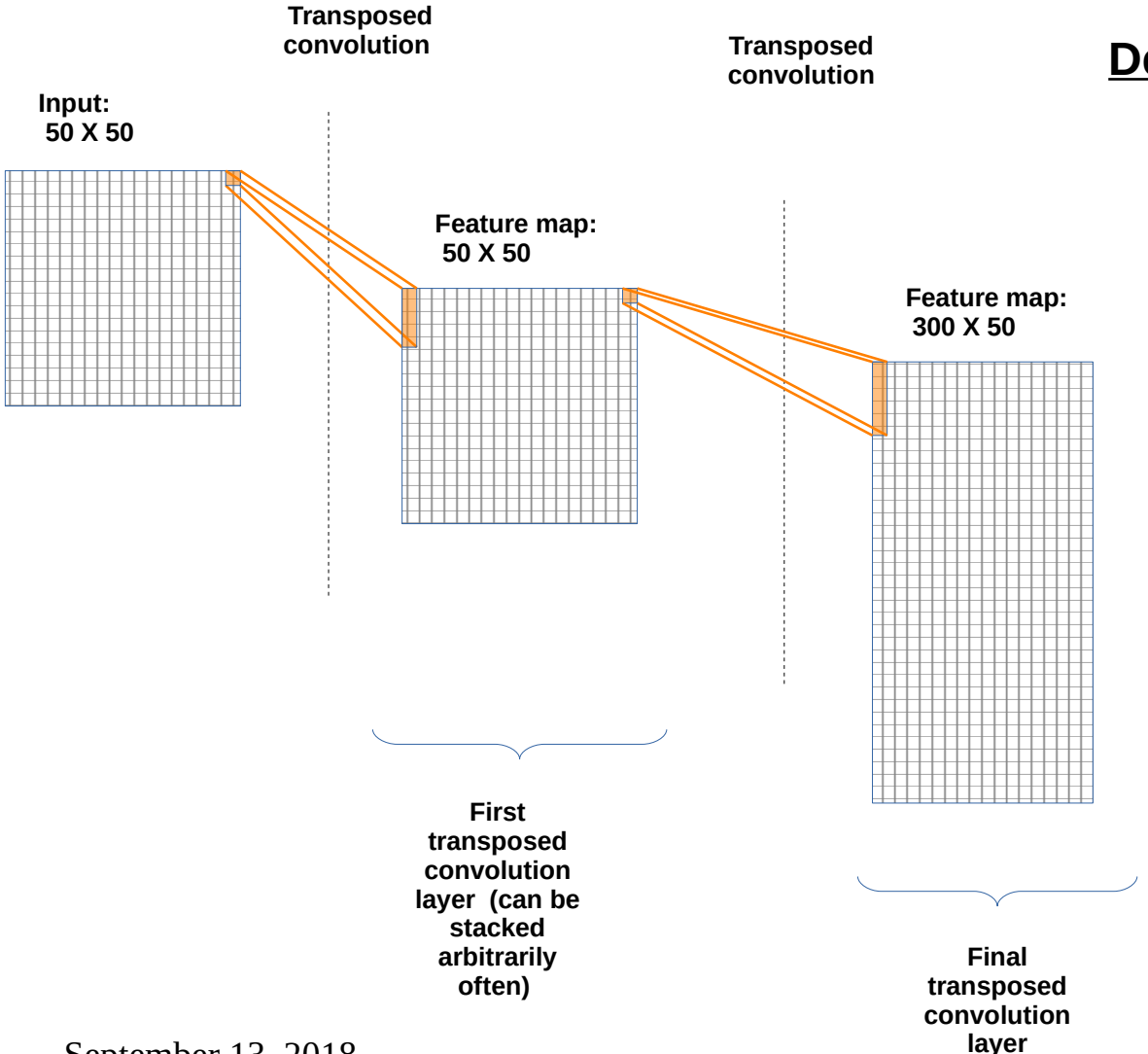
- Loss: normalized euclidean distance of feature map (50x50) of simple and complex sentence
- Layers: 2-4
- Best result: mean similarity of 60% with 4 Layers

$$\text{Similarity}(x_1,x_2) = 1-\text{loss}(x_1,x_2)$$

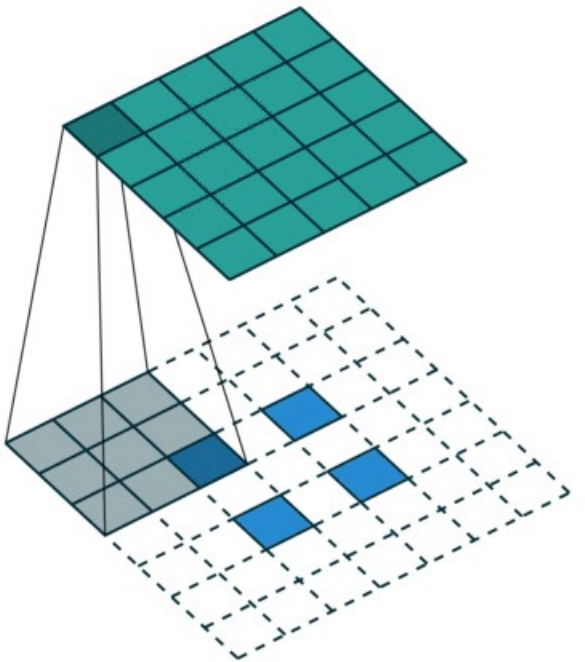
Encoder

- Switched to cosine similarity as loss function
  - Neglects magnitude of vectors but measures the orientation
  - FastText is trained with cosine similarity
- Layers: 4
- Result: mean similarity of 97.57% after 100 epochs and >98% after 1000 epochs

Decoder



Conv. Output = (i - k) + 2p + 1
Deconv. Output = (i' + k) – 2p -1



i=5,
k=4,
p=0,
i'=2

Image taken from: http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html

Decoder

- Loss: cosine similarity of final feature map to word embedding of simple sentence (both 300x50)
- Layers: 4 (same as Encoder)
- Result: mean similarity of 25%

Decoder

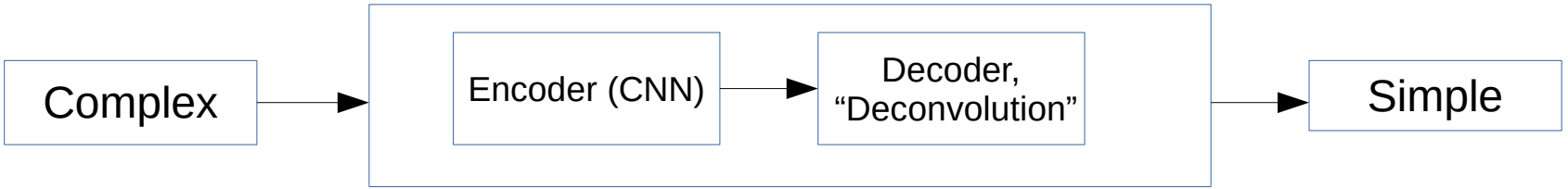
- Input:

The work of Müller and his wife with orphans began in 1836 with the preparation of their own rented home at 6 Wilson Street , Bristol for the accommodation of thirty girls .

- Output:

thảm da_id thảm cn/ pofi iró thảm dosimo dosimo dosimo dosimo thảm dosimo dosimo thảmالشباب←so ब्रिटिश #radioactive ಬೆಲ್ಲ sequestra cryolathe interposing toglie stereotypes_of_white_people liisi thảm dosimo dosimo dosimo dosimo dosimo dosimo dosimo dosimo dosimo dosimo thảm thảm

End-to-End network



- Loss: cosine similarity of feature map of complex sentence and word embedding of simple sentence
- Layers: 2-4
- Best result: mean similarity of 30% with 4 Layers

End-to-End network

- Input:

The work of Müller and his wife with orphans began in 1836 with the preparation of their own rented home at 6 Wilson Street , Bristol for the accommodation of thirty girls .

- Output:

錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 ‥ 錘 錘 ‥‥‥‥‥‥‥‥‥‥ 湊 *which which which which* 湊 湊 湊 湊 *bouïnian which garian kiraç seicho megalynarionkiraç mahārāshṭra* 堺 何鴻燊 *musuh aţ*

錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 錘 ・ 錘 錘 ‥ ɦ 錘 ‥ 湊 湊 ‥‥ɛ̋‥ 湊 ・ 湊 ・ 湊 湊 湊 湊 湊 *bouïnian which garian kiraç seicho megalynarion kiraç mahārāsh*

ṭra 堺 何鴻燊 *musuh aţ*
~

Conclusion

- Data set is noisy!
  - Additional information (co-reference)
  - Simple sentence 1.5 times the length of complex one
  - sometimes not really simplified matches
- Maybe train word embeddings on our corpus
- Encoder: very similar representation for simple and complex sentence
- Decoder: cannot reproduce word embedding of simple sentence from feature map of complex one
- End-to-end: cannot produce simple embedding from complex one