Convolutional Text Translation

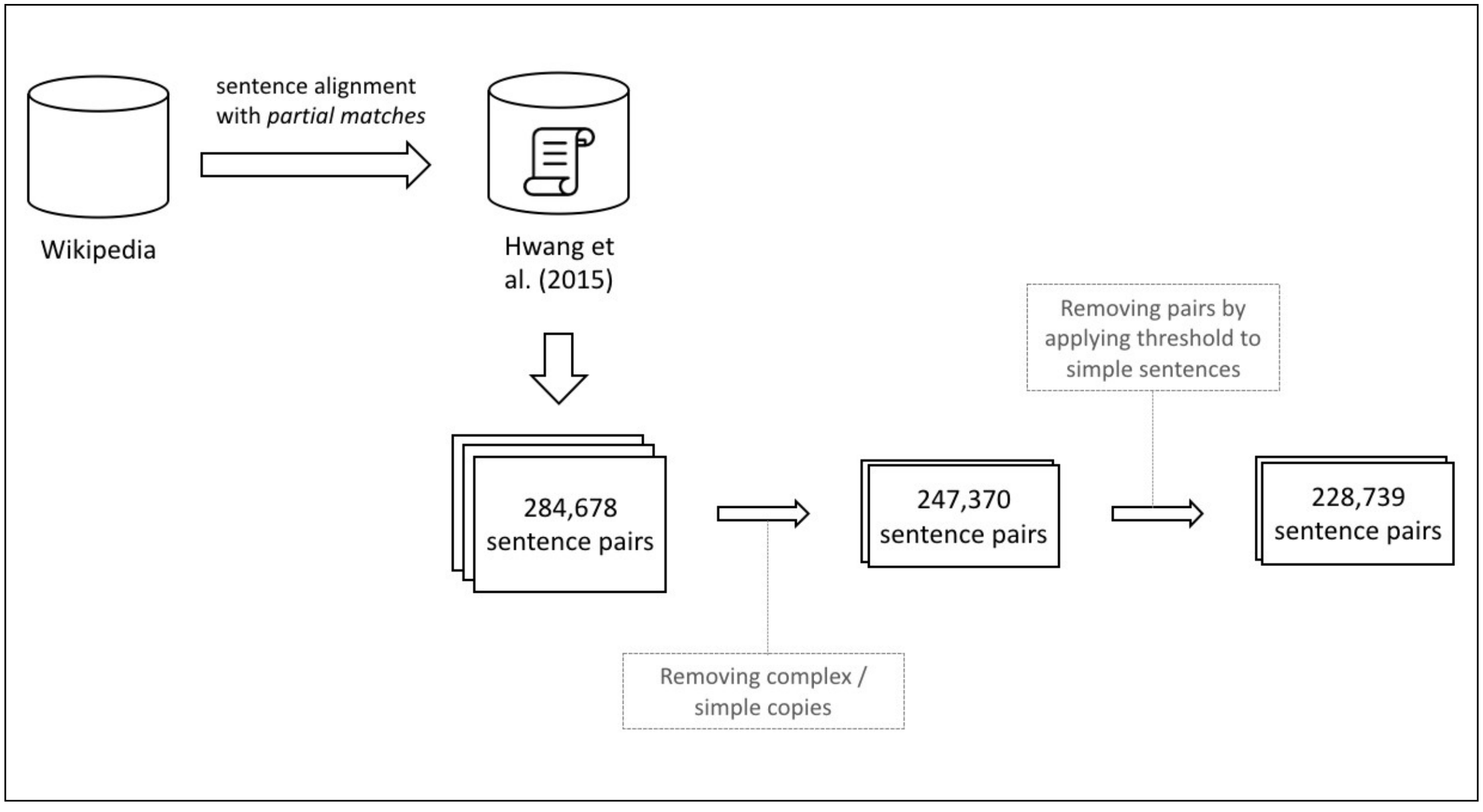
Simon Untergasser, Henny Sluyter-Gäthje,
Tim Patzelt, Dominik Pfütze

Why are we using a convolutional network?

CNN’s are state-of-the-art for Machine Translations:

by 2016 most of the best MT systems were using CNN’se.g. G*oogle, Microsoft or Yandex* translation services

Corpus


sentence alignment
with *partial matches*
Wikipedia
Hwang et
al. (2015)
284,678
sentence pairs
Removing complex /
simple copies
247,370
sentence pairs
Removing pairs by
applying threshold to
simple sentences
228,739
sentence pairs

Used Dataset

Hwang et al. (2015)

Match	Transformation	Sentence pair
Full	syntactic simplification; reordering of sentence constituents	"During the 13th century, gingerbread was brought to Sweden by German immigrants." and "German immigrants brought it to Sweden during the 13th century."
Full	lexical paraphrasing	"During the 13th century, gingerbread was brought to Sweden by German immigrants." and "German immigrants brought it to Sweden during the 13th century."
Partial	strong paraphrasing	"Gingerbread foods vary, ranging from a soft, moist loaf cake to something close to a ginger biscuit." and "Gingerbread is a word which describes different sweet food products from soft cakes to a ginger biscuit."
~~Partial~~	~~adding explanations~~	~~"Humidity is the amount of water vapor in the air." and "Humidity (adjective: humid) refers to water vapor in the air, but not to liquid droplets in fog, clouds, or rain."~~
Partial	sentence compression; dropping irrelevant information	"Falaj irrigation is an ancient system dating back thousands of years and is used widely in Oman, the UAE, China, Iran and other countries." and "The ancient falaj system of irrigation is still in use in some areas."

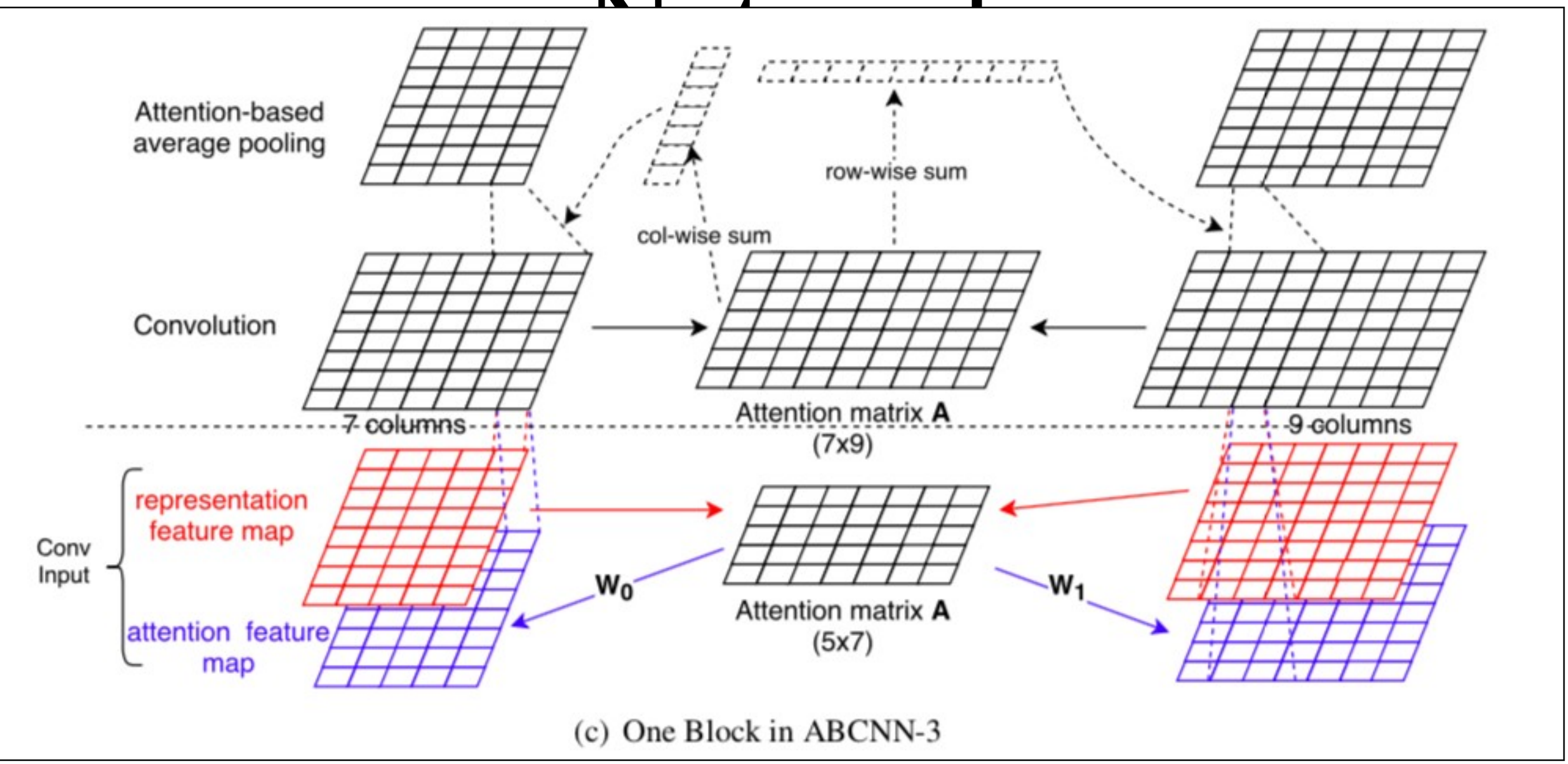
Why a Siamese Network?

- most prior work models each sentence separately
-

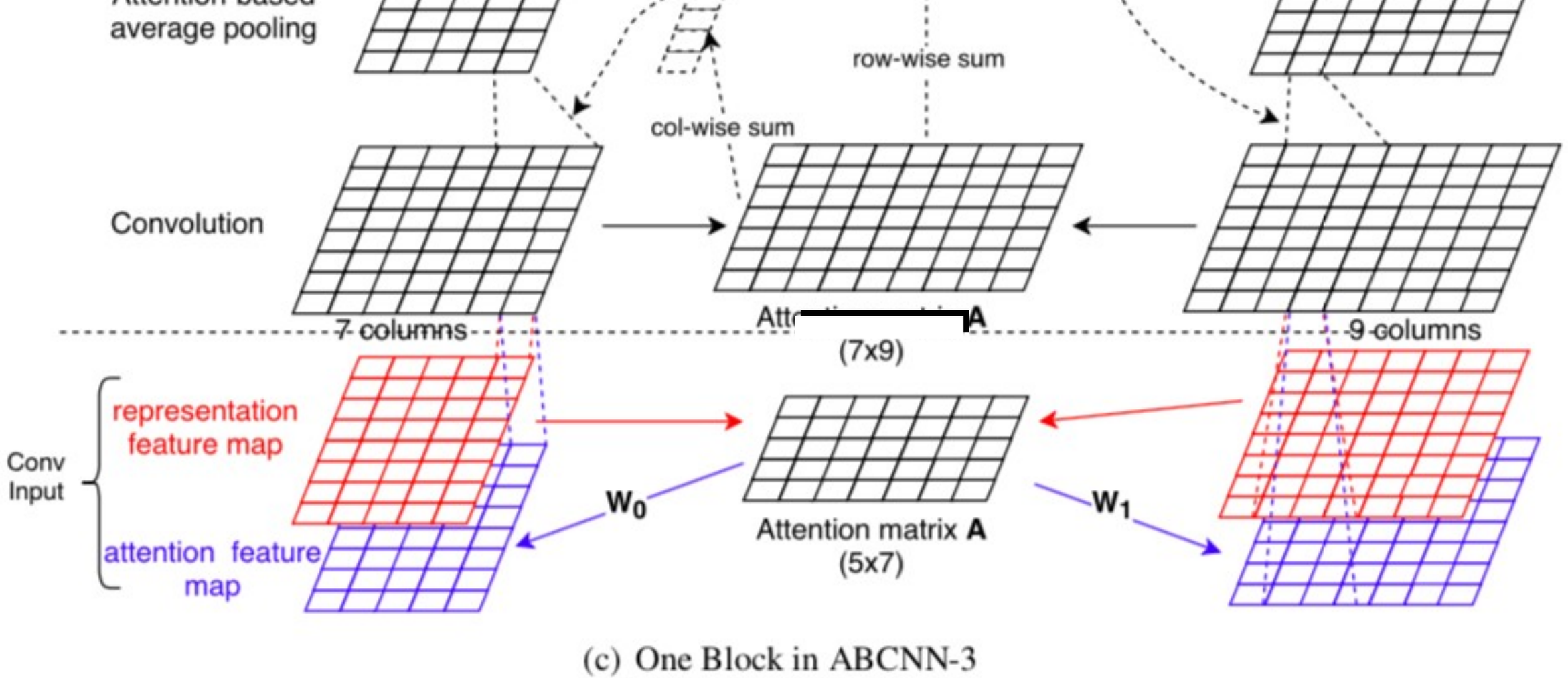
PI	$\mathbf{s_0}$	she struck a deal with RH to pen a book *today*
	$\mathbf{s_1^+}$	she signed a contract with RH to write a book
	$\mathbf{s_1^-}$	she denied *today* that she struck a deal with RH

- also contradicting to human behaviour:*we usually focus on key parts of one sentence by extracting parts from the other sentence that are related by identity, synonymy, antonymy and other relations*

Attention-Based Convolutional Neural Network



(c) One Block in ABCNN-3



(c) One Block in ABCNN-3

Given attention matrix $\mathbf{A}$, we generate the *attention feature map* $\mathbf{F}_{i,a}$ for s_i as follows:

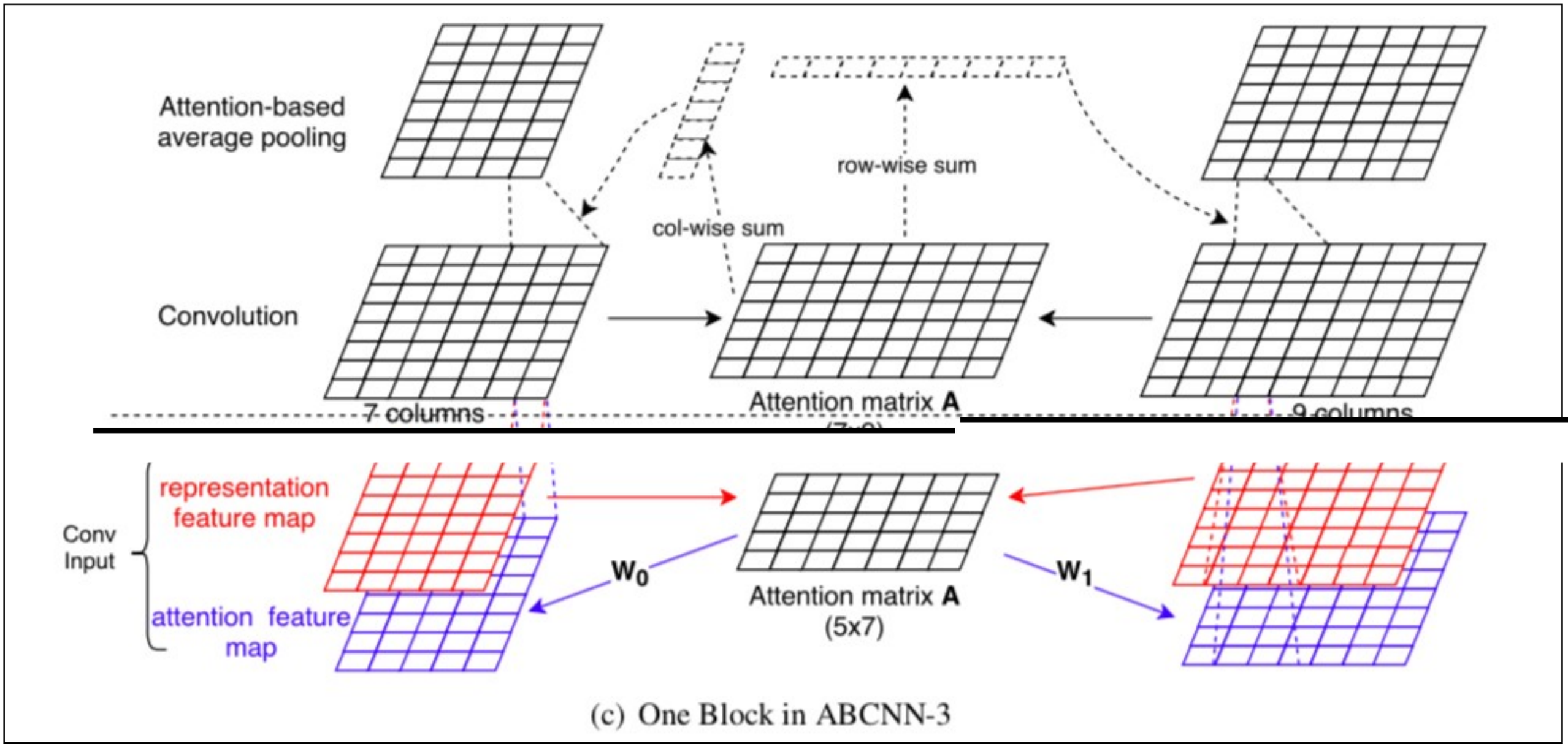
$$\mathbf{F}_{0,a} = \mathbf{W}_0 \cdot \mathbf{A}^\top \quad (3)$$

$$\mathbf{F}_{1,a} = \mathbf{W}_1 \cdot \mathbf{A} \quad (4)$$

The weight matrices $\mathbf{W}_0 \in \mathbf{R}^{d \times s}$, $\mathbf{W}_1 \in \mathbf{R}^{d \times s}$ are parameters of the model to be learned in training.

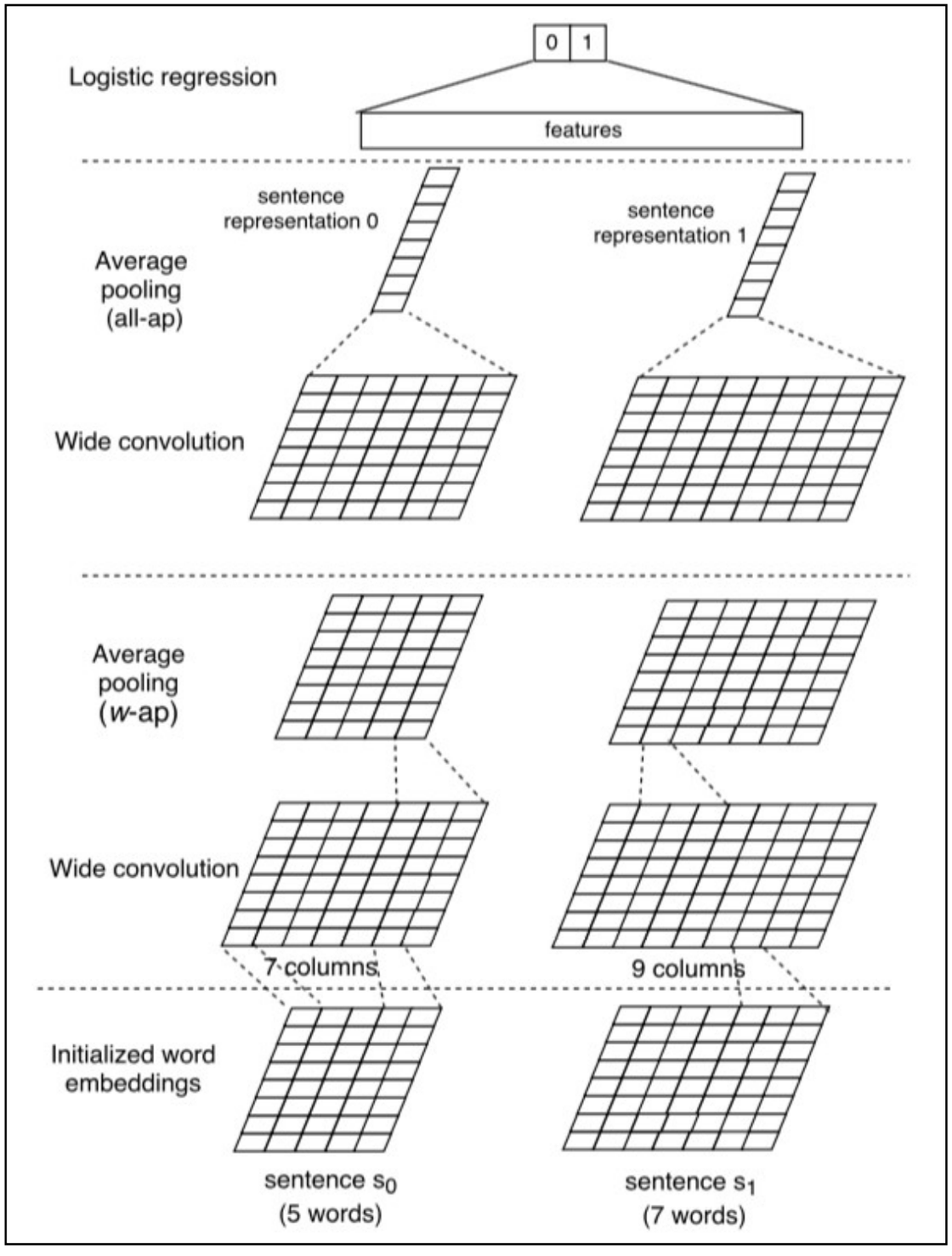
- computes attention weights on the *representation*
- *Motivation: the attention feature map shall guide the convolution to learn "counterpart-biased"*

Attention-based average pooling



(c) One Block in ABCNN-3

Full Net Architecture



Encoder

1:4 ratio of the input sentence:1 simple sentence matched correctly to the complex sentence4 simple sentences matched wrongly to the complex sentence

Decoder

Input:vector representation of complex sentence (Encoders' Output)

Output:the correct vector representation for the simple sentenceEvaluation:BLEU-score

Thank you