On the Benefit of Incorporating External Features in a Neural Architecture for Answer Sentence Selection

Ruey-Cheng Chen, Evi Yulianti, Mark Sanderson, W. Bruce Croft

RMIT UNIVERSITY

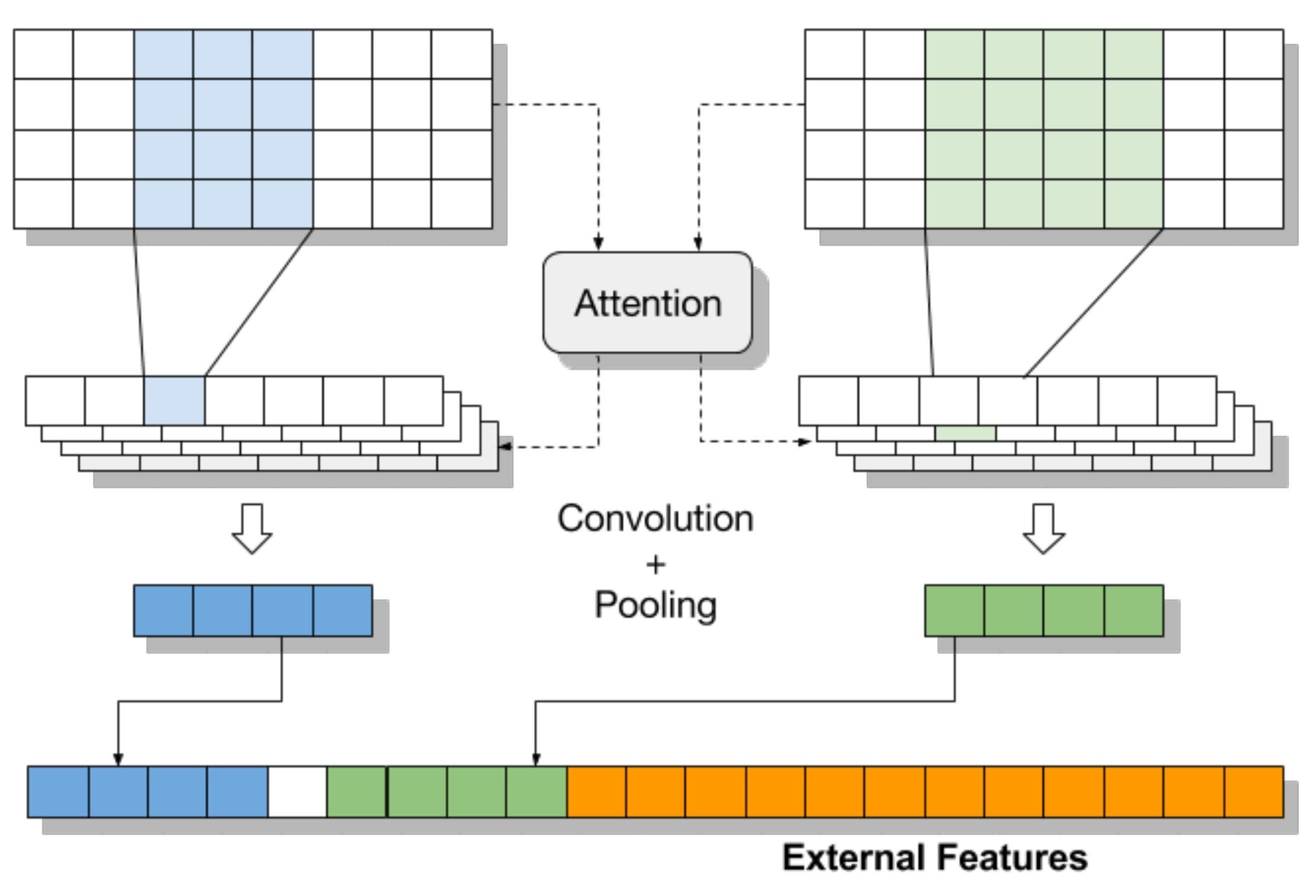
UMASS AMHERST

Does Deep Learning Remove the Need of Feature Engineering for Question Answering?

- Take any state-of-the-art neural question answering model
- Check if adding external features leads to further improvements
  - If yes, ignoring conventional features in evaluation makes *inaccurate performance assessments*.

Neural Network Configuration

- Bi-Convolutional Neural Networks (Severyn and Moschitti, 2015)
- Sparse word overlap indicators
- Kernel width 100; $\tanh$ activation; max pooling
- Batch size 50; AdaDelta trained; dev set based early stopping



External Features

- Lexical/semantic matching features (9)
- Readability features (8)
- Focus features (4)

Variables

- Word embeddings: Aquaint+wiki (50d), GoogleNews (300d)
- Dropout, swept through range $\{0.1, 0.2, \ldots, 0.9\}$
- Attention mechanism (ABCNN-1 model)

The attention layer takes question- and answer-side feature maps $\mathbf{F}_q \in \mathcal{R}^{n_q \times d}$ and $F_a \in \mathcal{R}^{n_a \times d}$ as input and computes $\mathbf{A} \in \mathcal{R}^{n_q \times n_a}$:

$$\mathbf{A}_{i,j} = \frac{1}{1 + \|\mathbf{F}_q[i,:] - \mathbf{F}_a[j,:]\|}, \qquad (1)$$

with $\|\cdot\|$ being the euclidean distance function. Two new attention-based feature maps, $\mathbf{F}'_q = AW_q$ and $\mathbf{F}'_a = A^T W_a$, are then to be combined in the follow-up convolutional layers.

Main Results

			TREC QA			WikiQA		
System	Attn?	Drop?	MAP	MRR	S@1	MAP	MRR	S@1
Runs (AQUAINT/Wikipedia)								
CNN	×	×	76.2	80.9	73.7	66.0	67.4	52.3
Combined Model	×	×	77.9 (+2.2%)	82.2 (+1.6%)	74.7 (+1.4%)	67.2 (+1.8%)‡	68.5 (+1.6%)‡	53.9 (+3.1%)‡
Combined Model	×	✓	**78.2 (+2.6%)**	**83.7 (+3.5%)**	76.8 (+4.2%)	64.7 (-2.0%)	65.7 (-2.5%)	48.6 (-7.1%)
CNN	✓	×	75.4	79.9	71.6	65.3	66.8	52.7
Combined Model	✓	×	77.2 (+2.4%)	81.1 (+1.5%)	72.6 (+1.4%)	70.0 (+7.2%)$^{\ddagger *}$	71.4 (+6.9%)$^{\ddagger *}$	58.4 (+10.8%)$^{\ddagger *}$
Combined Model	✓	✓	77.3 (+2.5%)	82.0 (+2.6%)	74.7 (+4.3%)	69.0 (+5.7%)‡	70.9 (+6.1%)$^{\ddagger *}$	58.4 (+10.8%)‡
Runs (Google News)								
CNN	×	×	76.1	82.3	75.8	67.3	69.1†	57.2‡
Combined Model	×	×	73.8 (-3.0%)	79.2 (-3.8%)	70.5 (-7.0%)	69.2 (+2.8%)‡	70.2 (+1.6%)‡	56.0 (-2.1%)‡
Combined Model	×	✓	74.8 (-1.7%)	80.1 (-2.7%)	71.6 (-5.5%)	69.2 (+2.8%)‡	70.7 (+2.3%)‡	56.4 (-1.4%)‡
CNN	✓	×	75.0	81.1	73.7	66.3	68.3	54.7‡
Combined Model	✓	×	76.5 (+2.0%)	82.5 (+1.7%)	74.7 (+1.4%)	69.4 (+4.7%)‡	71.2 (+4.2%)‡	57.6 (+5.3%)‡
Combined Model	✓	✓	76.3 (+1.7%)	82.5 (+1.7%)	74.7 (+1.4%)	67.9 (+2.4%)‡	69.7 (+2.0%)‡	56.0 (+2.4%)‡
Reference methods								
Bagged LambdaMART			75.7	81.3	72.6	63.0	63.8	46.5
LSTM (Wang et al., 2015)			71.3	79.1		—	—	
CNN (Severyn & Moschitti, 2015)			74.6	80.8		—	—	
aNMM (Yang et al., 2016)			75.0	81.1		—	—	
ABCNN-3 (Yin et al., 2015)			—	—		69.2	71.1	
PairwiseRank + SentLevel (Rao et al., 2016)			78.0	83.4		**70.1**	**71.8**	

Significant differences with respect to bagged LambdaMART and the group control are indicated by $^{\dagger}/^{\ddagger}$ and $^{*}/^{**}$, respectively, for $p < 0.05/p < 0.01$ using the paired t-test.