A Mention-Ranking Model for Abstract Anaphora Resolution – Supplementary Material –

Abstract

This document contains supplementary material to the paper "A Mention-Ranking Model for Abstract Anaphora Resolution".

1 Pre-processing details

The CSN corpus we obtained from the authors contained tokenized sentences for antecedents and anaphoric sentences. The number of instances differed from the reported numbers in KZH13 in 9 to 809 instances for training, and 1 for testing. The given sentences still contained the antecedent, so we removed it from the sentence and transformed the corresponding shell into "*this* ⟨*shell noun*⟩". An example of this process is: *The decision* ***to disconnect the ventilator*** *came after doctors found no brain activity.* → *This decision came after doctors found no brain activity.*

To use pre-trained word embeddings we had to lowercase all the data. As we use an automatic parse to extract all syntactic constituents, due to parser errors, candidates with the same string appeared with different tags. We eliminated duplicates by checking which tag is more frequent for candidates which have the same POS tag of the first word as the duplicated candidate, in the whole dataset. In case duplicated candidates were still occurring, we chose any of them. If such duplicates occur in antecedents, we don't take such instances in the training data to eliminate noise, or choose any of them for the test data. For the training data we choose instances with an anaphoric sentence length of at least 10 tokens.

All sentences in the batch are padded with a *PAD* token up to the maximal sentence length in the batch and corresponding hidden states in the LSTM are masked with zeros. To implement the model efficiently in TensorFlow, batches are constructed in such a way that every sentence instance in the batch has the same number of positive candidates and the same number of negative candidates. Note that by this we do **not** mean that the ratio of positive and negative examples is 1:1.

2 Hyperparameter details

Tables 1 and 2 report the tuned HPs for resolution of the shell noun *reason* and resolution of abstract anaphors in ARRAU-AA for different model variants. Below is the list of all tunable HPs.

- the dimensionality of the hidden states in the bi-LSTM, h_{LSTM}
- the first feed-forward layer size, h_{ffl1}
- the second feed-forward layer size, h_{ffl2}
- the dimensionality of the tag embeddings, d_{TAG}
- gradient clipping value, g
- frequency of words in vocabulary, f_w
- regularization coefficient, r
- keep probability of outputs of bi-LSTM, k_{LSTM}
- keep probability of input, k_{input}
- keep probability of outputs of the first feed-forward layer, k_{ffl1}
- keep probability of second of the first feed-forward layer, k_{ffl2}

We additionally report the number of trainable parameters (# param), the average epoch training time using one Nvidia GeForce GTX1080 gpu (t_e) and the epoch after which the best score is achieved (e).

ctx	aa	tag	cut	ffl1	ffl2	h_{LSTM}	h_{ffl1}	h_{ffl2}	d_{TAG}	g	f_w	r	k_{LSTM}	k_{ffl1}	k_{ffl2}	# param.	t_e	e
✓	✓	✓	✓	✓	✓	95	283	1115	49	2.13	9.40	6.61^{-5}	0.60	0.99	0.71	1928557	3.86	9
✗	✓	✓	✓	✓	✓	140	375	1193	83	7.44	4.41	2.87^{-6}	0.62	0.80	0.82	2842489	3.83	5
✓	✗	✓	✓	✓	✓	61	621	1485	81	8.27	3.27	3.44^{-3}	0.56	0.94	0.99	3502713	3.71	6
✓	✓	✗	✗	✓	✓	39	722	1655	-	43.00	7.11	2.91^{-6}	0.89	0.99	0.88	3624949	3.73	1
✓	✓	✓	✗	✓	✓	79	359	1454	65	3.22	7.74	9.66^{-6}	0.70	0.76	0.94	2459362	4.61	2
✗	✗	✗	✗	✓	✓	38	548	1997	-	82.00	7.14	6.07^{-3}	0.52	0.98	0.80	3345859	4.41	4
✓	✓	✓	✓	✗	✓	39	-	956	96	7.82	8.68	1.64^{-7}	0.78	-	0.59	1567647	4.41	3
✓	✓	✓	✓	✓	✗	71	305	-	94	9.40	5.42	8.3^{-3}	0.52	0.83	-	1593880	4.62	8

Table 1: HPs used for the different architecture variants for the shell noun *reason*.

ctx	aa	tag	cut	ffl1	ffl1	h_{LSTM}	h_{ffl1}	h_{ffl2}	d_{TAG}	g	f_w	r	k_{LSTM}	k_{input}	k_{ffl1}	k_{ffl2}	# param.	t_e	e
✓	✓	✓	✓	✓	✓	37	684	1081	99	7.40	2.41	2.38^{-4}	0.58	0.86	0.70	0.96	3716655	2.69	2
✗	✓	✓	✓	✓	✓	59	520	520	71	3.06	3.59	2.54^{-4}	0.70	0.83	0.56	0.89	2592937	2.62	1
✓	✗	✓	✓	✓	✓	45	782	447	31	1.50	6.20	1.22^{-4}	0.85	0.90	0.52	0.87	2300531	2.62	1
✓	✓	✗	✗	✓	✓	36	423	417	-	46.00	3.64	1.75^{-5}	0.57	0.86	0.65	0.63	2271652	8.24	2
✓	✓	✗	✗	✓	✓	36	423	417	-	46.00	3.64	1.75^{-5}	0.57	0.86	0.65	0.63	2271652	8.24	2
✓	✓	✗	✗	✓	✓	70	221	620	-	98	5.15	10^{-2}	0.90	0.87	0.84	0.75	2038202	8.26	1
✓	✓	✓	✗	✓	✓	121	355	1955	49	6.48	4.51	3.32^{-3}	0.87	0.90	0.77	0.84	3584370	8.33	1
✗	✗	✗	✗	✓	✓	44	622	633	-	96.00	4.29	1.49^{-5}	0.92	0.90	0.53	0.63	2541217	6.62	1
✓	✓	✓	✓	✗	✓	134	-	1489	36	9.51	2.44	3.87^{-3}	0.50	0.97	-	0.57	3575787	6.93	9
✓	✓	✓	✓	✓	✗	41	356	-	44	4.70	5.94	2.16^{-5}	0.66	0.94	0.97	-	1700229	2.64	1

Table 2: HPs used for evaluation on the ARRAU-AA test set.