Errata: Document Modeling with External Information for Sentence Extraction

Shashi Narayan
University of Edinburgh
shashi.narayan@ed.ac.uk

Ronald Cardenas
Charles University in Prague
ronald.cardenas@matfyz.cz

Nikos Papasarantopoulos
University of Edinburgh
nikos.papasa@ed.ac.uk

Shay B. Cohen **Mirella Lapata**
University of Edinburgh
{scohen,mlap}@inf.ed.ac.uk

Jiangsheng Yu **Yi Chang**
Huawei Technologies
{jiangsheng.yu,yi.chang}@huawei.com

Report on a Problem with the Evaluation in the Original Paper

In this errata we address an issue regarding the evaluation metrics used in our Answer Selection experiments (for the metrics ACC, MRR, and MAP).[1] Let s_i be the top ranked sentence in a document. Whenever s_i is not a correct answer, ACC gets a corresponding score of 0 added, whereas MRR has the value $\frac{1}{\text{rank}(s_i)}$ added to the total score. Hence, the ACC evaluation metric should always be smaller or equal than the MRR metric. This was not the case for our reported results.

Upon thorough inspection of the official TREC implementation[2] of MRR and MAP, we found out that ties (for the scores of sentences that are among the ones to be selected as an answer – the scores are based on the relevant model) are broken in such a way that the sentence that is picked is the one that comes first in inverse lexicographic order, treating the candidate sentence number id as a string (for example, according to that order, “100” is preferred over “2”). However, our implementation of the accuracy metric proposed by Trischler et al. (2016) breaks ties by choosing the candidate which comes earliest in the document (according to its index).

In order to remedy this inconsistency, we reimplemented all metrics with two tie-breaking options so that the setup can be consistent across metrics. Table 1 presents the results for the *first-in-line* setup, the case when ties are broken by choosing the candidate that comes *earliest* in the document. Likewise, Table 2 presents the results for the *last-in-line* setup, the case when ties are broken by choosing the candidate that comes *latest* in the document. The *last-in-line* implementation reproduces the results obtained with the official TREC scripts when leading zeros are added to the (string) ids of documents and candidates.[3]

The first observation to be mentioned about both Table 1 and 2 is that now ACC is smaller or equal to MRR in all cases. Second, it can observed that there is minimal variation of the results for the neural-based approaches when comparing both tie-breaking approaches. However, changes are significant for count-based baselines (WRDCNT, WGTWRDCNT; these methods are more likely to lead to a tie in scores for different sentences because they sum up scores for words that are in the intersection of the question and the candidate sentence. This set of words can be quite small, and as such there is less variability in these scores.

References

Cıcero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *CoRR, abs/1602.03609* .

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1400–1409. https://doi.org/10.18653/v1/D16-1147.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR* abs/1611.09830. http://arxiv.org/abs/1611.09830.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747* .

[1] We thank Javad Hosseini, Ming-Wei Chang and Kristina Toutanova for pointing out this issue.

[2] https://trec.nist.gov/trec_eval/

[3] Previous work, including in our own previous results, did not add add leading zeros to the ids under the (wrong) assumption of an integer sorting being internally performed by the evaluation script.

	SQuAD			WikiQA			NewsQA			MSMarco		
	ACC	MAP	MRR	ACC	MAP	MRR	ACC	MAP	MRR	ACC	MAP	MRR
Wrd Cnt	78.35	86.51	87.15	52.26	67.37	68.35	44.67	58.65	59.06	20.16	41.59	42.17
Wgt Wrd Cnt	78.94	86.56	87.27	51.44	66.91	67.48	45.24	58.43	58.86	20.50	41.85	42.43
LocalISF	79.99	87.55	88.22	49.79	66.3	66.99	44.69	58.36	58.73	20.21	41.78	42.31
ISF	79.35	86.81	87.52	51.03	66.56	67.21	45.61	58.74	59.16	20.52	41.86	42.43
PairCNN	33.05	55.63	55.76	30.86	50.11	51.10	22.83	38.09	38.33	14.28	35.17	35.81
CompAggr	85.88	91.04	91.79	61.32	72.76	73.70	54.52	67.61	68.19	32.05	52.82	53.43
XNet	36.86	59.09	59.44	54.73	68.28	69.30	26.19	42.70	42.85	15.45	36.66	37.25
XNetTopK	37.44	60.27	60.59	54.32	67.87	69.05	29.42	47.86	48.05	17.04	38.87	39.47
LRXNet	**85.98**	**91.13**	**91.88**	**63.37**	**74.71**	**75.40**	**58.84**	**72.71**	**73.09**	**32.93**	**53.41**	**54.03**
XNet+	79.83	87.35	88.04	55.56	68.89	70.06	47.26	61.58	61.97	23.07	44.95	44.38

Table 1: Results (in percentage) for answer selection using the *first-in-line* tie-breaking strategy, comparing the baselines (top) and our approaches (bottom).

	SQuAD			WikiQA			NewsQA			MSMarco		
	ACC	MAP	MRR	ACC	MAP	MRR	ACC	MAP	MRR	ACC	MAP	MRR
Wrd Cnt	77.61	85.48	86.25	29.63	49.26	49.58	31.69	44.13	44.54	20.61	41.98	42.59
Wgt Wrd Cnt	76.85	84.99	85.77	33.33	51.36	51.68	34.21	46.37	46.83	20.71	42.14	42.74
AP-CNN	-	-	-	-	68.86	69.57	-	-	-	-	-	-
ABCNN	-	-	-	-	69.21	71.08	-	-	-	-	-	-
L.D.C	-	-	-	-	70.58	72.26	-	-	-	-	-	-
KV-MemNN	-	-	-	-	70.69	72.65	-	-	-	-	-	-
LocalISF	78.87	86.41	87.16	31.28	50.37	50.85	34.31	46.67	47.11	20.64	42.09	42.65
ISF	77.25	85.23	86.01	32.92	50.96	51.44	34.72	46.77	47.20	20.70	42.13	42.73
PairCNN	21.50	46.23	46.26	18.11	39.42	40.25	22.83	38.09	38.32	13.94	34.96	35.54
CompAggr	85.88	91.04	91.79	61.32	72.76	73.70	54.52	67.61	68.19	32.08	52.84	53.45
XNet	33.90	56.62	56.74	54.73	68.28	69.30	25.93	44.34	44.67	13.81	34.93	35.5
XNetTopK	34.37	57.61	57.82	54.32	68.24	66.32	24.69	44.41	45.74	16.71	38.6	37.89
LRXNet	**85.98**	**91.13**	**91.88**	**62.96**	**74.29**	**74.98**	**58.84**	**72.71**	**73.09**	**32.93**	**53.41**	**54.03**
XNet+	78.96	86.58	87.32	55.56	68.89	70.06	39.16	53.18	53.53	18.50	38.98	39.65

Table 2: Results (in percentage) for answer selection using the *last-in-line* tie-breaking strategy, compared to previous work (top): AP-CNN (dos Santos et al., 2016), ABCNN (Yin et al., 2016), L.D.C (Wang and Jiang, 2016), KV-MemNN (Miller et al., 2016), and CompAggr Wang et al. (2017). The *last-in-line* setup is equivalent to the official TREC scripts when adding leading zeros to documents and candidates ids.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 189–198.

Wenpeng Yin, Hinrich Schtze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics* 4:259–272.