Supplementary Information for Continual Learning for Sentence Representations Using Conceptors

April 3, 2019

A The split STS datasets

In the main body of the paper, we have reported that we have used the STS datasets split by genre. A detailed list such STS tasks can be found in Table 1 and can be downloaded from `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark` and `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark#Companion`.

News	Captions	Forum	Tweets	WN
MSRpar 2012 headlines 2013-2016 deft-news 2014	MSRvid 2012 images 2014-2015 track5.en-en 2017	deft-forum 2014 answers-forums 2015 answer-answer 2016	tweet-news 2014	OnWN 2012-2014
4299 sentence pairs	3250 sentence pairs	1079 sentence pairs	750 sentence pairs	2061 sentence pairs

Table 1: STS datasets breakdown according to genres.

B CA compared with incremental-deletion SIF

We compare the CA approach with the following variant of SIF. In the learning phase, for each corpus coming in, we learn and store a common direction (estimated based on the new corpus). In the testing phase, for a sentence in the testing corpora, we project it away from all common directions we have stored so far. We call this approach SIF with incremental deletions. The testing result is reported in Figure 1.

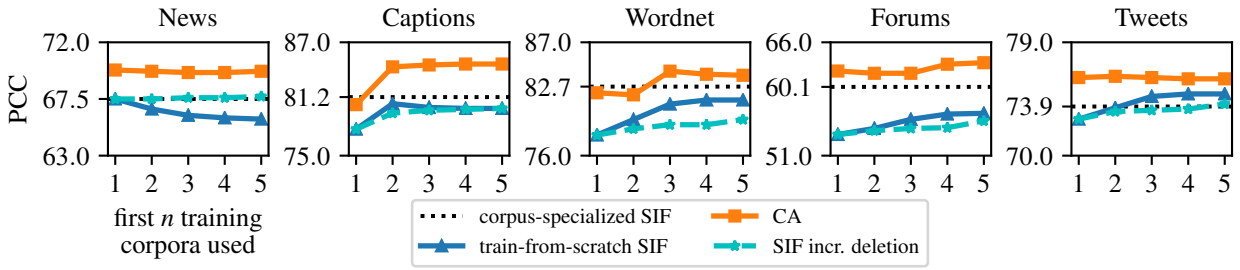

Figure 1: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. For explanation see text.

C CA without stop word initialization

We have also tested the performance of CA without the initializing our concepor C^0 by stop words. That is, we set C^0 as a zero matrix in our CA algorithm. The results are reported in Figure 2

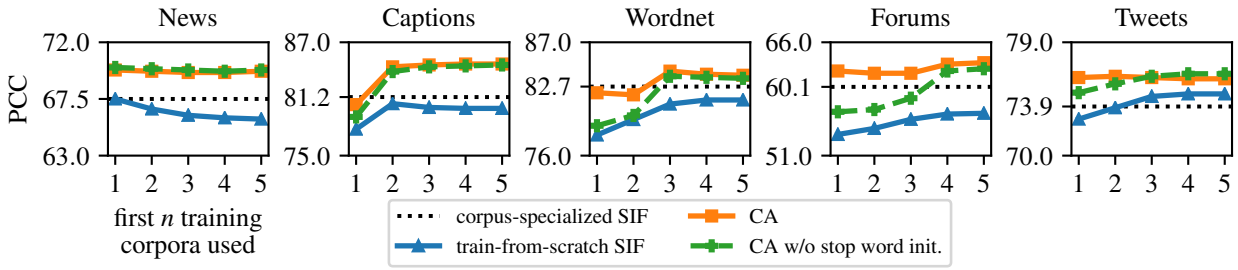

Figure 2: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. For explanation see text.

We see that, the CA initialized by stop words are more beneficial than without such initializations, especially for those testing corpora that are unseen in training data.

D CA with the reverse-ordered sequence of training corpora

In the main body of the paper, we sequentially presented new training corpus for sentence encoders, from the corpora of largest size (news) to that of the smallest size (tweets). We have remarked that this choice of ordering is essentially arbitrary. We now report the results for the reverse order (i.e., from corpora of smallest size to that of largest size) in Figure 3. We see that CA approach still outperforms train-from-scratch SIF throughout the time course.

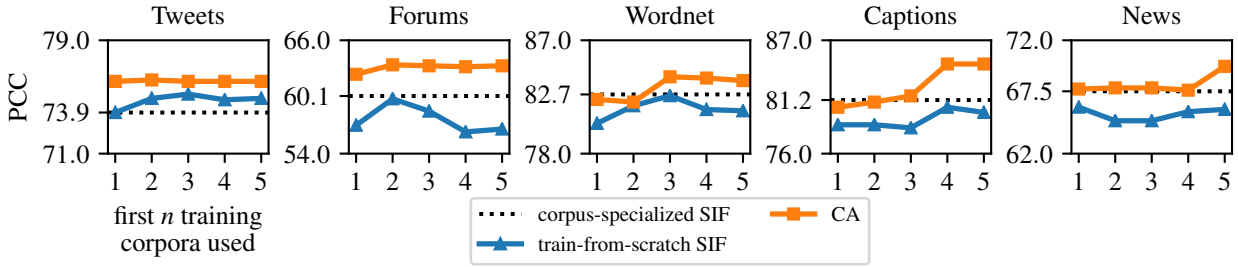

Figure 3: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. For explanation see text.

E Experiment using other word embedding brands

We repeat the experiments with Word2Vec [Mikolov et al., 2013][1] (pre-trained on Google News; 3 million tokens), Fasttext [Bojanowski et al., 2017] [2] (pre-trained on Common Crawl; 2 million of tokens), and Paragram SL-999[3] (fine-tuned based on GloVe). The pipeline of the experiments echo that of the main body of the paper.

[1] https://code.google.com/archive/p/word2vec/
[2] https://fasttext.cc/docs/en/english-vectors.html
[3] https://cogcomp.org/page/resource_view/106

E.1 Using Word2vec

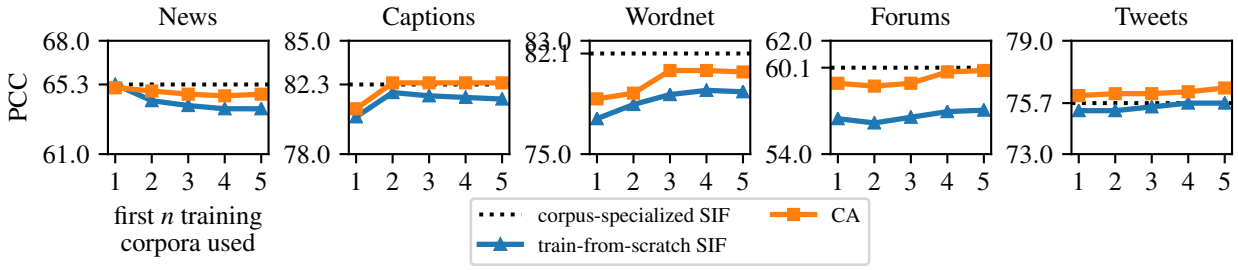

Figure 4: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. Word2Vec is used.

E.2 Using Fasttext

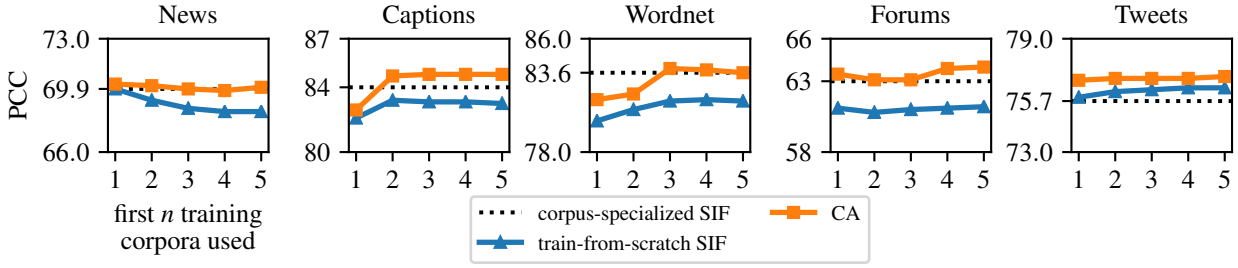

Figure 5: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. Fasttext is used.

E.3 Using Paragram-SL-999

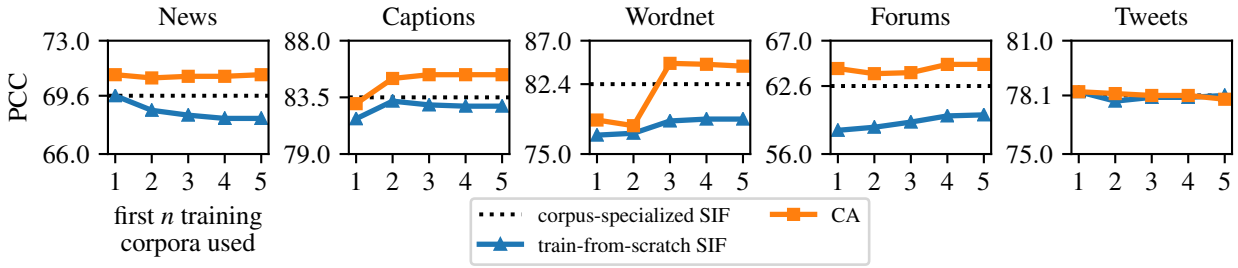

Figure 6: Pearson correlation coefficients (PCC) of the split STS datasets as a function of the number of training corpora. Paragram-SL-999 is used.

References

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.