Supplementary Material for CHARAGRAM: Embedding Words and Sentences via Character n-grams

John Wieting **Mohit Bansal** **Kevin Gimpel** **Karen Livescu**
Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
{jwieting,mbansal,kgimpel,klivescu}@ttic.edu

1 Training

For word and sentence similarity, we follow the training procedure of Wieting et al. (2015) and Wieting et al. (2016), described below. For part-of-speech tagging, we follow the English Penn Treebank training procedure of Ling et al. (2015).

For the similarity tasks, the training data consists of a set X of phrase pairs $\langle x_1, x_2\rangle$ from the Paraphrase Database (PPDB; Ganitkevitch et al., 2013), where x_1 and x_2 are assumed to be paraphrases. We optimize a margin-based loss:

$$\min_{\theta} \frac{1}{|X|}\Bigg(\sum_{\langle x_1,x_2\rangle \in X} \max(0, \delta - \cos(g(x_1), g(x_2)) + \cos(g(x_1), g(t_1))) + \max(0, \delta - \cos(g(x_1), g(x_2)) + \cos(g(x_2), g(t_2)))\Bigg) + \lambda \left\|\theta\right\|^2$$

where g is the embedding function in use, δ is the margin, the full set of parameters is contained in θ (e.g., for the CHARAGRAM model, $\theta = \langle W, \boldsymbol{b}\rangle$), λ is the L_2 regularization coefficient, and t_1 and t_2 are carefully selected negative examples taken from a mini-batch during optimization (discussed below). Intuitively, we want the two phrases to be more similar to each other ($\cos(g(x_1), g(x_2))$) than either is to their respective negative examples t_1 and t_2, by a margin of at least δ.

1.1 Selecting Negative Examples

To select t_1 and t_2 in Eq. 2, we tune the choice between two approaches. The first, MAX, simply chooses the most similar phrase in some set of phrases (other than those in the given phrase pair). For simplicity and to reduce the number of tunable parameters, we use the mini-batch for this set, but it could be a separate set. Formally, MAX corresponds to choosing t_1 for a given $\langle x_1, x_2\rangle$ as follows:

$$t_1 = \underset{t:\langle t,\cdot\rangle \in X_b \setminus \{\langle x_1,x_2\rangle\}}{\operatorname{argmax}} \cos(g(x_1), g(t))$$

where $X_b \subseteq X$ is the current mini-batch. That is, we want to choose a negative example t_i that is similar to x_i according to the current model parameters. The downside of this approach is that we may occasionally choose a phrase t_i that is actually a true paraphrase of x_i.

The second strategy selects negative examples using MAX with probability 0.5 and selects them randomly from the mini-batch otherwise. We call this sampling strategy MIX. We tune over the choice of strategy in our experiments.

2 Tuning Word Similarity Models

For all architectures, we tuned over the mini-batch size (25 or 50) and the type of sampling used (MIX or MAX). δ was set to 0.4 and the dimensionality d of each model was set to 300.

For the CHARAGRAM model, we tuned the activation function h ($\tanh$ or linear) and regularization coefficient λ (over $\{10^{-4}, 10^{-5}, 10^{-6}\}$). The n-gram vocabulary V contained all 100,283 character n-grams ($n \in \{2, 3, 4\}$) in the lexical section of PPDB XXL.

For charCNN and charLSTM, we randomly initialized 300 dimensional character embeddings for

Dataset	50%	75%	Max	charCNN	charLSTM	PARAGRAM-PHRASE	CHARAGRAM-PHRASE
MSRpar	51.5	57.6	73.4	50.6	23.6	42.9	**59.7**
MSRvid	75.5	80.3	88.0	72.2	47.2	76.1	**79.6**
SMT-eur	44.4	48.1	56.7	50.9	38.5	45.5	**57.2**
OnWN	60.8	65.9	72.7	61.8	53.0	**70.7**	68.7
SMT-news	40.1	45.4	60.9	46.8	38.3	57.2	**65.2**
STS 2012 Average	54.5	59.5	70.3	56.5	40.1	58.5	**66.1**
headline	64.0	68.3	78.4	68.1	54.4	72.3	**75.0**
OnWN	52.8	64.8	84.3	54.4	33.5	**70.5**	67.8
FNWN	32.7	38.1	58.2	26.4	10.6	**47.5**	42.3
SMT	31.8	34.6	40.4	42.0	24.2	40.3	**43.6**
STS 2013 Average	45.3	51.4	65.3	47.7	30.7	**57.7**	57.2
deft forum	36.6	46.8	53.1	45.6	19.4	50.2	**62.7**
deft news	66.2	74.0	78.5	73.5	54.6	73.2	**77.0**
headline	67.1	75.4	78.4	67.4	53.7	69.1	**74.3**
images	75.6	79.0	83.4	68.7	53.6	**80.0**	77.6
OnWN	78.0	81.1	87.5	66.8	46.1	**79.9**	77.0
tweet news	64.7	72.2	79.2	66.2	53.6	76.8	**79.1**
STS 2014 Average	64.7	71.4	76.7	64.7	46.8	71.5	**74.7**
answers-forums	61.3	68.2	73.9	47.2	27.3	**67.4**	61.5
answers-students	67.6	73.6	78.8	75.0	63.1	78.3	**78.5**
belief	67.7	72.2	77.2	65.7	22.6	76.0	**77.2**
headline	74.2	80.8	84.2	72.2	61.7	74.5	**78.7**
images	80.4	84.3	87.1	70.0	52.8	82.2	**84.4**
STS 2015 Average	70.2	75.8	80.2	66.0	45.5	75.7	**76.1**
2014 SICK	71.4	79.9	82.8	62.9	50.3	**72.0**	70.0
2015 Twitter	49.9	52.5	61.9	48.6	39.9	52.7	**53.6**
Average	59.7	65.6	73.6	59.2	41.9	66.2	**68.7**

Table 1: Results on SemEval textual similarity datasets (Pearson's $r \times 100$). The highest score in each row is in boldface (omitting the official task score columns).

all unique characters in the training data. For charLSTM, we tuned over whether to include an output gate. For charCNN, we tuned the filter activation function (rectified linear or $\tanh$) and tuned the activation for the fully-connected layer ($\tanh$ or linear). For both the charLSTM and charCNN models, we tuned λ over $\{10^{-4}, 10^{-5}, 10^{-6}\}$.

3 Full Sentence Embedding Results

Table 1 shows the full results of our sentence similarity experiments.

References

[Ganitkevitch et al.2013] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of HLT-NAACL*.

[Ling et al.2015] Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

[Wieting et al.2015] John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL (TACL)*.

[Wieting et al.2016] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of International Conference on Learning Representations*.