Learning sentence representations from NLI data

Xinyi Chen
University of Amsterdam
xinyi.chen@student.uva.nl

1 Quantitative Analysis

		NLI		Transfer	
Model	**dim**	**dev**	**test**	**micro**	**macro**
AWE	300	60.55	61.02	79.94	77.31
LSTM	2048	80.47	80.55	80.6	78.37
BiLSTM	4096	78.84	79.65	81.96	80.35
BiLSTM-Max	4096	83.9	84.12	84.03	82.68

Table 1: Performance of sentence encoder architectures on SNLI and (aggregated) transfer tasks.

2 Qualitative Analysis

Changing word orders Changing word orders makes no difference to the AWE encoder outputs, but changes values in almost all dimensions of LSTM and BLSTM encoder outputs. When the word order changes do not lead to different sentence meanings, values in most dimensions of BLSTM-Max encoder outputs remain the same. But they vary when changing word orders modify meanings. In such cases, only the classifier with BLSTM-Max encoding makes correct predictions.

Replacing non-content words BLSTM-Max encoder neglects non-content word changes when the sentence meaning doesn't change. Values in most embedding dimensions remain the same. But the other encoders are sensitive to these changes. One interesting finding is that when changing prepositions to opposite meanings(e.g."in" to "outside"), the values in many dimensions do have minor change, but the classifier still makes wrong predictions. In such case, the classifier with AWE encoding performs better.

Changing tenses In error analysis examples, the change of tense doesn't make much difference of BLSTM-Max encoder outputs, which lead to wrong predictions in NLI tasks. For BLSTM and LSTM encoders, the values in most embedding dimension are different, but classifiers fail to make correct predictions. The AWE encoder have different outputs and is able to predict correct labels in some cases.

Changing verb object The classifiers with LSTM, BLSTM and BLSTM-max encoding make wrong predictions for a pair that changes verb object. The BLSTM-max encoding represent the two sentences with very similar word embeddings, which might suggest that its failure in distinguishing the differences. In this case, using AWE encoding helps the classifier make the correct prediction.

3 Summary

Overall, the BLSTM-max encoders outperforms the other models in all tasks and more robust to changes in sentences. But it has some limitations. Though classifier with AWE encoding performs worse, it performs better than other models in some cases.

A Appendix

	MR	**CR**	**SUBJ**	**MPQA**	**SST2**	**TREC**	**MRPC**	**SICK-R**	**SICK-E**	**STS14**
AWE	77.46	79.13	91.18	87.52	80.07	82.8	71.88	75.50	73.01	.55/.55
LSTM	74.0	78.22	86.71	87.85	77.81	76.2	73.8	85.67	82.42	.61/.62
BiLSTM	75.3	79.1	90.14	88.01	79.57	86.4	74.32	86.42	84.49	.60/.62
BiLSTM-Max	77.94	80.88	92.23	88.67	82.32	89.6	75.07	88.33	85.18	.67/.68

Table 2: Transfer test results for various architectures trained in different ways.