WOLVESAAR at SemEval-2016 Task 1: Replicating the Success of Monolingual Word Alignment and Neural Embeddings for Semantic Textual Similarity

Hanna Bechara, Rohit Gupta, Liling Tan,
Constantin Orăsan, Ruslan Mitkov, Josef van Genabith
University of Wolverhampton / UK,
Universität des Saarlandes / Germany
Deutsches Forschungszentrum für Künstliche Intelligenz / Germany
{hanna.bechara, r.gupta, corsasan, r.mitkov}@wlv.ac.uk,
liling.tan@uni-saarland.de, josef.van_genabith@dfki.de

Abstract

This paper describes the WOLVESAAR systems that participated in the English Semantic Textual Similarity (STS) task in SemEval-2016. We replicated the top systems from the last two editions of the STS task and extended the model using GloVe word embeddings and dense vector space LSTM based sentence representations. We compared the difference in performance of the replicated system and the extended variants. Our variants to the replicated system show improved correlation scores and all of our submissions outperform the median scores from all participating systems.

1 Introduction

Semantic Textual Similarity (STS) is the task of assigning a real number score to quantify the semantic likeness of two text snippets. Similarity measures play a crucial role in various areas of text processing and translation technologies ranging from improving information retrieval rankings (Lin and Hovy, 2003; Corley and Mihalcea, 2005) and text summarization to machine translation evaluation and enhancing matches in translation memory and terminologies (Resnik and others, 1999; Ma et al., 2011; Banchs et al., 2015; Vela and Tan, 2015).

The annual SemEval STS task (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015) provides a platform where systems are evaluated on the same data and evaluation criteria.

2 DLS System from STS 2014 and 2015

For the past two editions of the STS task, the top performing submissions are from the DLS@CU team (Sultan et al., 2014b; Sultan et al., 2015).

Their STS2014 submission is based on the proportion of overlapping content words between the two sentences treating semantic similarity as a monotonically increasing function of the degree to which two sentences contain semantically similar units and these units occur in similar semantic contexts (Sultan et al., 2014b). Essentially, their semantic metric is based on the proportion of aligned content words between two sentences, formally defined as:

$$prop_{Al}^{(1)} = \frac{|\{i : [\exists j : (i,j) \in Al] \; and \; w_i^{(1)} \in C\}|}{|\{i : w_i^{(1)} \in C\}|} \quad (1)$$

where $prop_{Al}^{(1)}$ is the monotonic proportion of the semantic unit alignment from a set of alignments Al that maps the positions of the words (i, j) between sentences $S^{(1)}$ and $S^{(2)}$, given that the aligned units belong to a set of content words, C. Since the proportion is monotonic, the equation above only provides the proportion of semantic unit alignments for $S^{(1)}$. The Al alignments pairs are automatically annotated by a monolingual word aligner (Sultan et al., 2014a) that uses word similarity measures based on contextual evidence from the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) and syntactic dependencies.

The same computation needs to be made for $S^{(2)}$. An easier formulation of the equation without the formal logic symbols is:

Proceedings of SemEval-2016, pages 634–639,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

$$prop_{Al}^{(1)} = \frac{sum(1\ for\ w_i, w_j\ in\ Al^{(1,2)}\ if\ w_i\ in\ C)}{sum(1\ for\ w_i\ in\ S^{(1)}\ if\ w_i\ in\ C)} \quad (2)$$

Since the semantic similarity between $(S^{(1)}, S^{(2)})$ should be a single real number, Sultan et al. (2014b) combined the proportions using harmonic mean:

$$sim(S^{(1)}, S^{(2)}) = \frac{2 * prop_{Al}^{(1)} * prop_{Al}^{(2)}}{prop_{Al}^{(1)} + prop_{Al}^{(2)}} \quad (3)$$

Instead of simply using the alignment proportions, Sultan et al. (2015) extended their hypothesis by leveraging pre-trained neural net embeddings (Baroni et al., 2014). They posited that the semantics of the sentence can be captured by the centroid of its content words[1] computed by the element-wise sum of the content word embeddings normalized by the number of content words in the sentence. Together with the similarity scores from Equation 3 and the cosine similarity between two sentence embeddings, they trained a Bayesian ridge regressor to learn the similarity scores between text snippets.

3 Our Replica of DLS for STS 2016

To replicate the success of Sultan et al. (2014b), we use the monolingual word aligner from Sultan et al. (2014a) to annotate the STS-2012 to STS-2015 datasets and computed the alignment proportions as in Equation 1 and 2.

In duplicating Sultan et al. (2015) work, we first have to tokenize and lemmatize text. The details of pre-processing choices was undocumented in their paper, thus we lemmatized the datasets with the NLTK tokenizer (Bird et al., 2009) and PyWSD lemmatizer (Tan, 2014). We use the lemmas to retrieve the word embeddings from the COMPOSES vector space (Baroni et al., 2014). Similar to Equation 2 (changing only the numerator), we sum the sentence embedding's centroid as follows:

$$v(S^{(1)}) = \frac{sum(v(w_i)\ for\ w_i\ in\ S^{(1)}\ if\ w_i\ in\ C)}{sum(1\ for\ w_i\ in\ S^{(1)}\ if\ w_i\ in\ C)} \quad (4)$$

where $v(S^{(1)})$ refers to the dense vector space representation of the sentence $S^{(1)}$ and $v(w_i)$ refers to the word embedding of word i provided by the COMPOSES vector space. The same computation has to be done for $S^{(2)}$.

Intuitively, if either of the sentences contains more or less content words than the other, we can see the numerator changing but the denominator changes with it. The difference between $v(S^{(1)})$ and $v(S^{(2)})$ contributes to *distributional semantic distance*.

To calculate a real value similarity score between the sentence vectors, we take the dot product between the vectors to compute the cosine similarity between the sentence vectors:

$$sim(S^{(1)}, S^{(2)}) = \frac{v(S^{(1)}) \cdot v(S^{(2)})}{|v(S^{(1)})|\ |v(S^{(2)})|} \quad (5)$$

There was no clear indication of which vector space Sultan et al. (2015) have chosen to compute the similarity score from Equation 5. Thus we compute two similarity scores using both COMPOSES vector spaces trained with these configurations:

- 5-word context window, 10 negative samples, subsampling, 400 dimensions
- 2-word context window, PMI weighting, no compression, 300K dimensions

In this case, we extracted two similarity features for every sentence pair. With the harmonic proportion feature from Equation 3 and the similarity scores from Equation 5, we trained a boosted tree ensemble on the 3 features using the STS 2012 to 2015 datasets and submitted the outputs from this model as our baseline submission in the English STS Task in SemEval 2016.

3.1 Replacing COMPOSES with GloVe

Pennington et al. (2014) handles semantic regularities (Levy et al., 2014) explicitly by using a global log-bilinear regression model which combines the global matrix factorization and the local context vectors when training word embeddings.

Instead of using the COMPOSES vector space, we experimented with replacing the $v(w_i)$ com-

[1] In the implementation, they have used lemmas instead of words to reduce sparsity when looking up the pre-trained embeddings (personal communication with Arafat Sultan).

ponent in Equation 4 with the GloVe vectors,[2] $v_{glove}(w_i)$ such that:

$$sim_{glove}(S^{(1)}, S^{(2)}) = \frac{v_{glove}(S^{(1)}) \cdot v_{glove}(S^{(2)})}{|v_{glove}(S^{(1)})|\;|v_{glove}(S^{(2)})|} \tag{6}$$

The novelty lies in the usage of the global matrix to capture corpus wide phenomena that might not be captured by the local context window. The model leverages on both the non-zero elements in the word-word co-occurence matrix (not a sparse bag-of-words matrix) and the individual context window vectors similar to the word2vec model (Mikolov et al., 2013).

3.2 Similarity Using Tree LSTM

Recurrent Neural Nets (RNNs) allow arbitrarily sized sentence lengths (Elman, 1990) but early work on RNNs suffered from the vanishing/exploding gradients problem (Bengio et al., 1994). Hochreiter and Schmidhuber (1997) introduced multiplicative input and output gate units to solve the vanishing gradients problem. While RNN and LSTM process sentences in a sequential manner, Tree-LSTM extends the LSTM architecture by processing the input sentence through a syntactic structure of the sentence. We use the ReVal metric (Gupta et al., 2015) implementation of Tree-LSTM (Tai et al., 2015) to generate the similarity score.

ReVal represents both sentences (h_1, h_2) using Tree-LSTMs and predicts a similarity score $\hat{y}$ based on a neural network which considers both distance and angle between h_1 and h_2:

$$\begin{aligned} h_{\times} &= h_1 \odot h_2 \\ h_{+} &= |h_1 - h_2| \\ h_s &= \sigma\left(W^{(\times)} h_{\times} + W^{(+)} h_{+} + b^{(h)}\right) \\ \hat{p}_\theta &= \text{softmax}\left(W^{(p)} h_s + b^{(p)}\right) \\ \hat{y} &= r^T \hat{p}_\theta \end{aligned} \tag{7}$$

where, σ is a sigmoid function, $\hat{p}_\theta$ is the estimated probability distribution vector and $r^T = [1\ 2 ... K]$. The cost function $J(\theta)$ is defined over probability distributions p and $\hat{p}_\theta$ using regularised Kullback-Leibler (KL) divergence.

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{KL}\left(p^{(i)} \middle\| \hat{p}_\theta^{(i)}\right) + \frac{\lambda}{2} ||\theta||_2^2 \tag{8}$$

In Equation 8, i represents the index of each training pair, n is the number of training pairs and p is the sparse target distribution such that $y = r^T p$ is defined as follows:

$$p_j = \begin{cases} y - \lfloor y \rfloor, & j = \lfloor y \rfloor + 1 \\ \lfloor y \rfloor - y + 1, & j = \lfloor y \rfloor \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq j \leq K$, where, $y \in [1, K]$ is the similarity score of a training pair. This gives us a similarity score between [1, K] which is mapped between [0, 1].[3] Please refer to Gupta et al. (2015) for training details.

4 Submission

We submitted three models based on the original replication of the Sultan et al. (2014b) and Sultan et al. (2015) system and our variants and extensions of their approach.

Our `baseline` submission uses the similarity score from Equations 3 and 5 as features to train a linear ridge regression. Our `baseline` submission achieved an overall 0.69244 Pearson correlation score on all domains.

Extending the `baseline` implementation, we included the similarity score from Equations 6 and 8 to the feature set and trained a boosted tree ensemble (Friedman, 2001) to produce our `Boosted` submission. Finally, we use the same feature set to train an eXtreme Boosted tree ensemble (XGBoost) (Chen and He, 2015; Chen and Guestrin, 2015) model.

We annotated the STS 2012 to 2015 datasets with the similarity scores from Equations 2, 3, 5, 6, 8. The annotations and our open source implementation of the system are available at `https://github.com/alvations/stasis/blob/master/notebooks/STRIKE.ipynb`

[2] We use the 300 dimensions vectors from the GloVe model trained on the Commoncrawl Corpus with 840B tokens, 2.2M vocabulary.

[3] score = (score-1)/K

	answer-answer	headlines	plagiarism	postediting	question-question	All
Baseline	0.48799	0.71043	0.80605	0.84601	0.61515	0.69244
Boosted	0.49415	0.71439	**0.79655**	0.83758	**0.63509**	0.69453
XGBoost	**0.49947**	**0.72410**	0.79076	**0.84093**	0.62055	**0.69471**
+Saarsheff	0.50628	0.77824	0.82501	0.84861	0.70424	0.73050
Median	0.48018	0.76439	0.78949	0.81241	0.57140	0.68923
Best	0.69235	0.82749	0.84138	0.8669	0.74705	0.77807

Table 1: Pearson Correlation Results for English STS Task at SemEval-2016

5 Results

Table 1 shows the results of our submission to the English STS task in SemEval-2016; the median and best scores are computed across all participating teams in the task. Our baseline system performs reasonably well, outperforming the median scores in most domains.

Our extended variant of the baseline using boosted tree ensemble performs better in the answer-answer, headlines and postediting domains but performed worse in others. Comparatively, it improves the overall correlation score marginally by 0.002.

The system using XGBoost performs the best of the 3 models but it underperforms in the headlines and plagiarism domain when compared to the median scores.

Generally, we did not achieve the outstanding scores in the task as compared to the top performing team DLS@CU in the English STS 2015. Our XGBoost system performs far from the best scores from the top systems. However, overall our correlation scores are higher than the median scores across all submissions for the task.

As a post-hoc test, we have evaluated our baseline system by training on the STS 2012 to 2014 dataset and testing on the STS 2015 dataset and we achieved 0.76141 weighted mean Pearson correlation score on all domains. As compared to Sultan et al. (2015) results of 0.8015 we are 0.04 points short of their results which should technically rank our system at 20th out of 70+ submissions to the STS 2015 task[4].

Machine Translation (MT) evaluation metrics have shown competitive performance in previous STS tasks (Barrón-Cedeño et al., 2013; Huang and Chang, 2014; Bertero and Fung, 2015; Tan et al., 2015). Tan et al. (2016) annotated the STS datasets with MT metrics scores for every pair of sentence in the training and evaluation data. We extend our XGBoost model with these MT metric annotations and achieved a higher score for every domain leading to an overall Pearson correlation score of 0.73050 (+Saarsheff in Table 1).

6 Conclusion

In this paper, we have presented our findings on replicating the top system in the STS 2014 and 2015 task and evaluated our replica of the system in the English STS task of SemEval-2016. We have introduced variants and extensions to the replica system by using various state-of-art word and sentence embeddings. Our systems trained on (eXtreme) Boosted Tree ensembles outperform the replica system using linear regression. Although our replica of the previous best system did not achieve stellar scores, all our systems outperform the median scores computed across all participating systems.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, pages 385–393, Montréal, Canada.

[4] Our replication attempt obtained better results compared to our STS-2015 submission (MiniExperts) that used a Support Vector Machine regressor trained on a number of linguistically motivated features (Gupta et al., 2014); it achieved 0.7216 mean score (Béchara et al., 2015).

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 32–43, Atlanta, Georgia.

Eneko Agirre, Carmen Banea, Claire Cardic, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado.

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(3):472–482.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland.

Alberto Barrón-Cedeño, Lluís Màrquez, Maria Fuentes, Horacio Rodríguez, and Jordi Turmo. 2013. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 143–147, Atlanta, Georgia.

Hanna Béchara, Hernani Costa, Shiva Taslimipoora, Rohit Guptaa, Constantin Orăsan, Gloria Corpas Pastorb, and Ruslan Mitkova. 2015. Miniexperts: An svm approach for measuring semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 96–101.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166.

Dario Bertero and Pascale Fung. 2015. Hltc-hkust: A neural network paraphrase classifier using translation metrics, semantic roles and lexical similarity features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 23–28, Denver, Colorado.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python.* " O'Reilly Media, Inc.".

Tianqi Chen and Carlos Guestrin. 2015. Xgboost: Reliable large-scale tree boosting system.

Tianqi Chen and Tong He. 2015. xgboost: extreme gradient boosting. *R package version 0.4-2*.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia.

Rohit Gupta, Hanna Béchara, Ismail El Maarouf, and Constantin Orăsan. 2014. Uow: Nlp techniques developed at the university of wolverhampton for semantic similarity and textual entailment. In *8th Int. Workshop on Semantic Evaluation (SemEval14)*, pages 785–789.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Pingping Huang and Baobao Chang. 2014. SSMT:A Machine Translation Evaluation View To Paragraph-to-Sentence Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 585–589, Dublin, Ireland.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning: a translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1239–1248.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Philip Resnik et al. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. Dls@ cu: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. Dls@ cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 148–153.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China.

Liling Tan, Carolina Scarton, Lucia Specia, and Josef van Genabith. 2015. Usaar-sheffield: Semantic textual similarity with deep regression and machine translation evaluation metrics. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 85–89, Denver, Colorado.

Liling Tan, Carolina Scarton, Lucia Specia, and Josef van Genabith. 2016. Saarsheff at semeval-2016 task 1: Semantic textual similarity with machine translation evaluation metrics and (extreme) boosted tree ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California.

Liling Tan. 2014. Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd.

Mihaela Vela and Liling Tan. 2015. Predicting machine translation adequacy with document embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 402–410, Lisbon, Portugal.