DalGTM at SemEval-2016 Task 1: Importance-Aware Compositional Approach to Short Text Similarity

Jie Mei, Aminul Islam, Evangelos Milios
Faculty of Computer Science
Dalhousie University, Canada
{jmei,islam,eem}@cs.dal.ca

Abstract

This paper describes our system submission to the SemEval 2016 English Semantic Textual Similarity (STS) shared task. The proposed system is based on the compositional text similarity model, which aggregates pairwise word similarities for computing the semantic similarity between texts. In addition, our system combines word importance and word similarity to build an importance-similarity matrix. Three different word similarity measures are used in our three submitted runs. The evaluation results show that taking into account context dependent word importance information improves performance. However, the performance of the system varies drastically between different evaluation subsets. The best of our submitted runs achieves rank 60th with weighted mean Pearson correlation to human judgements of 0.6892.

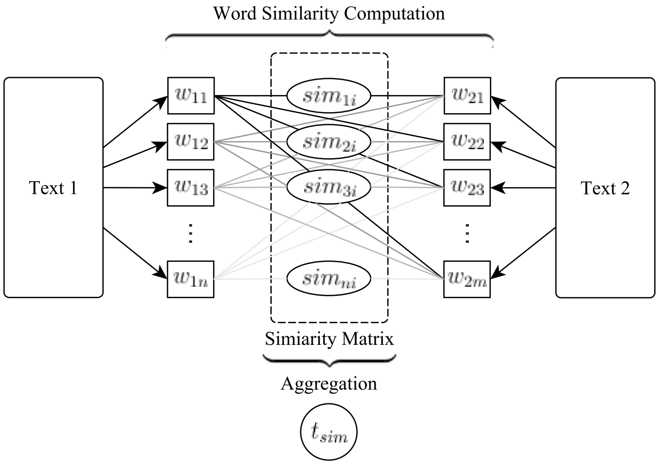

Figure 1: The general procedure of the compositional text similarity measures. They take three general steps: tokenize the input text, compute pairwise word similarities between all words, and aggregate the resulting scores to a sentence level textual similarity score. $\{w_{11}, w_{12}, \ldots, w_{1n}\}$ and $\{w_{21}, w_{22}, \ldots, w_{1m}\}$ are the tokenized words from Text 1 and Text 2, respectively. Each node in the middle represents a vector of pairwise similarity values computed by one word from Text 1 and all distinct words from Text 2.

1 Introduction

Semantic Textual Similarity (STS) measures the degree of equivalence in the underlying semantics of paired natural language texts. It is an extensively researched problem with applications widely used in many research areas including natural language processing, information retrieval, and text mining. The STS task has been held annually since 2012 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirrea et al., 2015) to encourage research into understanding sentence-level semantics. Systems for this task compute semantic similarity scores for paired text snippets. Performance is evaluated by the Pearson correlation between the system scores and human judgements.

This paper describes our system submission to the SemEval 2016 STS shared task (Agirre et al., 2016). The proposed system is based on the compositional text similarity model, which have been broadly researched in the literature by (Mihalcea et al., 2006; Li et al., 2006; Islam and Inkpen, 2006; Ho et al., 2010; Islam et al., 2012; Bär, 2013). The compositional text similarity model makes use of word-level

Proceedings of SemEval-2016, pages 765–770,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

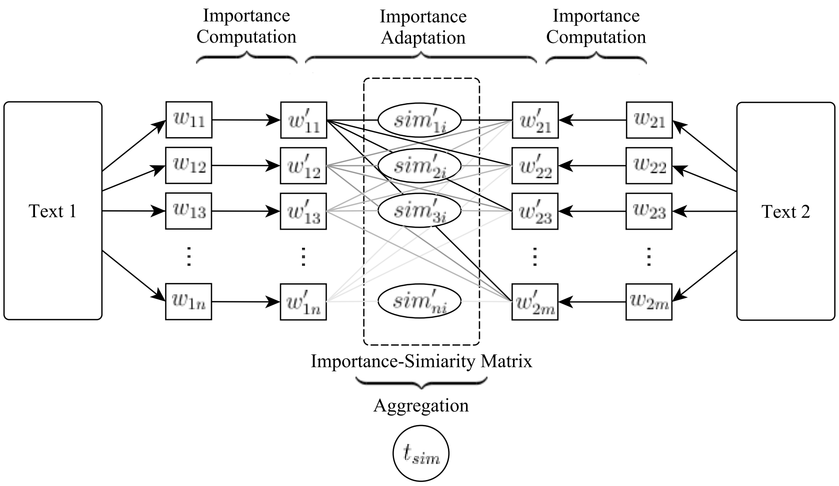

Figure 2: The procedure of the proposed importance-aware similarity measure. The general compositional procedure in Fig. 1 uses context independent similarity value. In addition, this measure takes into account context dependent word importance information. It first computes the word importance value w' using Eq. 3 and adapted into every entry sim' in the importance-similarity matrix using Eq. 4.

similarity values as the building blocks to compute sentence-level semantics. Computing textual similarity using this approach proceeds as follows: tokenize the input text, compute pairwise word similarities between all words, and aggregate the resulting scores to a sentence level textual similarity score. State-of-the-art word similarity measures can be used in this model to provide context independent word relatedness. However, words an be more or less important depending on the contexts in which they appear.

We extend traditional compositional models with an importance term for each word. Our three submitted runs use this extended model in combination with three different word similarity measures: Google Trigram Method (Islam et al., 2012), Skip-grams word embedding (Mikolov et al., 2013), and GloVe word embedding (Pennington et al., 2014). The evaluation results show that including matching importance information improves the performance of compositional models on most of the evaluation sets for STS 2015 and 2016. However, the relative performance of our systems varies dramatically when comparing against other systems submitted to the shared task. The best of our submitted runs achieves rank 60th with weighted mean Pearson correlation of 0.6892 with human judgements.

The rest of the paper is organized as follows: Section 2 describes the details of the submitted systems. Section 3 shows the experimental results for our three runs using evaluation data from SemEval 2015 and 2016. Section 4 summarizes our observations and concludes.

2 System Description

Our proposed approach takes advantage of the compositional model while also taking the importance of words into consideration. It first computes the matching importance, which characterizes the importance of a word in a particular text pair similarity computation. Instead of building the similarity matrix used by the traditional compositional approach (Fig. 1), we construct an alternative importance weighted similarity matrix. The importance weighted similarity values are then used when computing the overall textual similarity score.

2.1 Text Preprocessing

The input texts are tokenized using the Penn Treebank tokenization with additional rules from Google.[1] Punctuation and the 33 most common English words are filtered because these tokens contribute little to the semantic meaning of a text. Then, we lemmatized the remaining words taking into account their POS tags. The preprocessed input text

[1] Decribed in Section 2.2 on https://catalog.ldc.upenn.edu/docs/LDC2006T13/readme.txt

is represented by these lemmas. POS tagging and lemmatization use NLTK toolkit.[2]

2.2 Word Similarity Computation

Word similarities are the core building blocks in compositional text similarity measures. Our three different runs each explore using a different word similarity algorithm.

Google Trigram Method (Islam et al., 2012) is an unsupervised statistical similarity measure that can be applied to word pairs. This word similarity method characterizes the co-occurrence feature using the frequencies of trigrams starting and ending with a word pair. It is computed using the following formula:

$$\varphi'(w_1, w_2) = \frac{\frac{\frac{1}{2}(f_n(w_1,w_2)+f_n(w_2,w_1))}{\min(f(w_1),f(w_2))}}{\frac{f(w_1)}{f_{\max}} \cdot \frac{f(w_2)}{f_{\max}}}, \quad (1)$$

where $f_n(w_1, w_2)$ indicates the total frequency of n-grams starting with w_1 and ending with w_2. $f(w)$ stands for the word (i.e. uni-gram) frequency of w and $f_{\max}$ is the largest word frequency in the corpus. Then, the following normalization function is applied to Eq. 1 to bound the word similarity values in range $[0, 1]$:

$$sim(w_1, w_2) = \begin{cases} \frac{\log \varphi'(w_1,w_2)}{-2\times\log \frac{\min(f(w_1),f(w_2))}{C_{\max}}} & \text{if } \log \varphi'(w_1,w_2)>1 \\ \frac{\log 1.01}{-2\times\log \frac{\min(f(w_1),f(w_2))}{f_{\max}}} & \text{if } \log \varphi'(w_1,w_2)\leq 1 \\ 0 & \text{if } f_n(w_1,w_2)+f_n(w_2,w_1)=0. \end{cases} \quad (2)$$

We use the efficient implementation of this method described in Mei et al. (2015).

Skip-grams (Mikolov et al., 2013) is a neural network model for learning word embeddings. Word embeddings are trained using a model that attempts to discriminatively predict word co-occurrences within a fixed context window. The resulting word embedding vectors have been shown to be effective at capturing word-level semantic information. We use the pre-trained vectors that were learned on a part of Google News dataset (about 100 billion words).[3]

GloVe (Pennington et al., 2014) is an unsupervised learning algorithm for word embeddings. The method learns word embedding vectors using a model that predicts global word co-occurrence statistics extracted from a corpus. We use the pre-trained vectors built using the Wikipedia 2014 dump and the English Gigaword Fifth Edition.[4]

For Skip-gram and GloVe, we use cosine similarity to compute pairwise similarity value.

2.3 Matching Importance Computation

We define matching importance as a function that characterizes the importance of a word in a particular textual similarity computation. Given w in one text and $w_1, w_2, \ldots, w_n$ in the other text, the matching importance of w is computed by this expression:

$$imp(w) = \alpha \cdot \mu(S) + \beta \cdot \sigma(S) \quad (3)$$

$$s.t. \quad S = \{s \mid s = sim(w, w_i), 1 \leq i \leq n\},$$

where function μ and ρ stands for the mean and standard deviation of a set of values. This expression is used in Islam et al. (2012) for selecting important matchings. The mean of similarities is an indicator of semantic relatedness, whereas the standard deviation indicates distinctiveness. We take the weighted sum of both features as the final importance score. In our system submissions, we set $\alpha = \beta = 1$.

2.4 Matching Importance Adaptation

To incorporate the importance information, we rescale the pairwise word-level similarity scores by the minimum of the context dependent importance scores for the words being compared:

$$sim'(w_i, w_j) = sim(w_i, w_j) \cdot \min(imp(w_i), imp(w_j)). \quad (4)$$

2.5 Textual Similarity Computation

Given a preprocessed text pair, we count (δ) and remove the identical words in both texts. Let the

[2] http://www.nltk.org/

[3] https://code.google.com/archive/p/word2vec/

[4] http://nlp.stanford.edu/projects/glove/

Year	Subset Name	Description	#Pairs
2015	answer-forums	forums answers	375
	answer-students	student short answers	750
	belief	belief annotations	750
	headline	news headlines	750
	images	image descriptions	750
2016	answer-answer	stackexchange answers	254
	headline	news headline	249
	plagiarism	plagiarised short answers	230
	question-question	stackexchange questions	209
	postediting	machine translations with post-editions	244

Table 1: A brief description of SemEval 2015 and 2016 datasets. The SemEval 2015 and 2016 datasets contain test sentence pairs distributed across nine domains.

remaining words be $T_1 = \{w_{11}, \ldots, w_{1n}\}$ and $T_2 = \{w_{21}, \ldots, w_{2m}\}$, we construct an importance weighted similarity matrix $M_{n \times m}$. Prior work suggests that only using the most important entries in the matrix may suppress interference during semantic analysis. (Mihalcea et al., 2006; Islam and Inkpen, 2008; Islam et al., 2012) Thus, we set up a threshold t_i to filter the less important matchings in the ith row of the matrix:

$$t_i = \mu(S') + \rho(S') \tag{5}$$

$$s.t. \quad S' = \{s \mid s = sim'(w_{1i}, w_{2j}), 1 \leq j \leq m\}.$$

The textual similarity between two sentences is computed as follows in Eq. 6:

$$t_{sim} = \frac{(\delta + \sum_{1 \leq i \leq n} \mu(S))(n + m + 2\delta)}{2(n + \delta)(m + \delta)} \tag{6}$$

$$\begin{aligned} s.t. \quad S = \{s \mid s = sim'(w_{1i}, w_{2j}), s \geq t_i, \\ 1 \leq i \leq n, 1 \leq j \leq m\}. \end{aligned}$$

sim' is the importance weighted similarity from Eq. 4. $n + \delta$ and $m + \delta$ are the lengths of two preprocessed texts. The textual similarity score ranges within $[0, 1]$.

3 Evaluation

We evaluated our three system submissions using the STS 2015 and 2016 evaluation datasets. The SemEval 2015 and 2016 datasets contain test sentence pairs distributed across nine domains. Each pair was assigned a similarity scores in the range [0, 5] by multiple human annotators. The performance of our three system submissions is shown in Table 2 and 3. Recall that our three systems only differ in the method they use for assessing lexical similarity: Google Trigram Method (*GTM*), Word2vec (*W2V*), and *GloVe*. Systems that make use of matching importance are tagged with *+IAC*. Otherwise, the system directly uses pairwise similarity values to compute the aggregate similarity score using Eq. 6. Note that systems with the proposed matching importance approach perform consistently better than the original compositional model in most of the domain subsets. This shows that adding an importance feature can effectively improve the performance of the compositional model. However, comparing against the average system performance in each domain, the performance of our submitted systems vary dramatically in their relative performance to systems submitted by other participating teams. For example, our systems perform well on the postediting dataset and dramatically worse, even relative to other systems, on the question-question data. This suggests that the proposed system may have an implicit domain specific bias.

4 Conclusions and Future Work

In this paper, we present an Importance-Aware Compositional Approach to STS and its evaluation during the SemEval 2016 STS shared task. Experimental results show that the proposed approach performs consistently better than matched compositional similarity models that do not take importance into account. In future work, it would be useful to investigate a more robust weighting scheme for word importance, incorporating syntactic analysis of texts and using external knowledge-bases for word sense disambiguation.

Acknowledgments

The research was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and The Boeing Company.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot

system	answer-forums	answer-students	belief	headline	images	average
GloVe	0.4220	0.6959	0.5901	0.6998	0.7775	0.6371
GloVe+IAC	0.5301	0.7134	0.6893	0.6992	0.7928	0.6850
GTM	0.5963	0.7021	0.6837	0.6880	0.7676	0.6875
GTM+IAC	0.6065	0.7128	0.7203	0.6940	0.7841	0.7035
W2V	0.6033	0.7164	0.7181	0.6939	0.7932	0.7050
W2V+IAC	0.6116	0.7336	0.6961	0.7033	0.8150	0.7119
average official	0.5705	0.6599	0.6382	0.7220	0.7689	0.6888
best official	0.7390	0.7725	0.7491	0.8250	0.8644	0.8015

Table 2: Evaluation result for SemEval 2015 STS dataset. Both the compositional model and the proposed model (systems with *+IAC*) are implemented with three word similarity measures: *GTM*, *W2V*, and *GloVe*. In most of the comparison experiments, the proposed model gets higher Pearson correlation than the original compositional model with the same setting. However, the performance of our submitted systems varies dramatically comparing with the average system performance in different domain subsets.

system	answer-answer	headlines	plagiarism	question-question	postediting	average
GTM	0.4894	0.6717	0.7791	0.4271	0.8525	0.6440
GTM+IAC	0.5137	0.6907	0.7969	0.4331	0.8478	0.6564
W2V	0.5136	0.6791	0.8174	0.4739	0.8522	0.6672
W2V+IAC	0.5365	0.6855	0.8087	0.4710	0.8456	0.6695
GloVe	0.4745	0.6957	0.7994	0.5065	0.8301	0.6721
GloVe+IAC	0.5285	0.6961	0.7994	0.5480	0.8301	0.6804
average official	0.4802	0.7644	0.7895	0.5714	0.8124	0.6892
best official	0.6924	0.8275	0.8414	0.7471	0.8669	0.7781

Table 3: Evaluation results for SemEval 2016 STS dataset. It shows the same characteristics of the proposed model in both SemEval datasets.

on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385–393, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-agirre, and Weiwei Guo. 2013. sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*.

Eneko Agirre, Carmen Banea, Claire Cardie, 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.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity - monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, USA, June. Association for Computational Linguistics.

Eneko Agirrea, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Weiwei Guof, Inigo Lopez-Gazpioa, Montse Maritxalara, Rada Mihalceab, et al. 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.

Daniel Bär. 2013. *A Composite Model for Computing Similarity Between Texts*. Ph.D. thesis, TU Darmstadt, TU Darmstadt, 10.

Chukfong Ho, Masrah Azrifah Azmi Murad, Rabiah Abdul Kadir, and Shyamala C. Doraisamy. 2010. Word sense disambiguation-based sentence similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 418–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aminul Islam and Diana Inkpen. 2006. Second order co-occurrence pmi for determining the semantic similarity of words. In *Proceedings of the 21st International Conference on Language Resources and Evaluation*, pages 1033–1038, Genoa, Italy.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2(2):10:1–10:25, July.

Aminul Islam, Evangelos Milios, and Vlado Kešelj. 2012. Text similarity using google tri-grams. In *Advances in Artificial Intelligence*, pages 312–317. Springer Berlin Heidelberg.

Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, Aug.

Jie Mei, Xinxin Kou, Zhimin Yao, Andrew Rau-Chaplin, Aminul Islam, Abidalrahman Moh'd, and Evangelos E. Milios. 2015. Efficient computation of co-occurrence based word relatedness. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 43–46, New York, NY, USA. ACM.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

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