Using Semantic Information Space to Evaluate Semantic Textual Similarity

Hao Wu, Heyan Huang , Ping Jian, Yuhang Guo, Chao Su

BIT at SemEval-2017 (Task 1)

Problems Addressed

Model

This paper presents three systems for semantic textual similarity (STS).

- One is an unsupervised system and the other two are supervised systems which simply employ the unsupervised one.
- All the systems depend on the semantic information space (SIS), which is constructed based on the semantic hierarchical taxonomy in WordNet.
- The systems try to compute non-overlapping information content (IC) of sentences.

Proposed Solution

Preliminaries

1. Information content IC of concept c whose statistical frequency is $P(c)$

$$IC(c) = -\log P(c)$$

2. Similarity of two sentences S_a and S_b on the basis of Jaccard coefficient is given by:

$$sim(s_a, s_b) = \frac{IC(s_a \cap s_b)}{IC(s_a \cup s_b)}$$

   Where quantity of intersection can be calculated by:

$$IC(s_a \cap s_b) = IC(s_a) + IC(s_b) - IC(s_a \cup s_b)$$

3. Information content of sentences can be calculated by inclusion-exclusion principle, as:

$$IC(s_a) = IC(\bigcup_{i=1}^{n} c_i^a) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < ... < i_k \leq n} IC(c_{i_1}^a \cap ... \cap c_{i_k}^a)$$

4. Common IC of concepts is given by:

$$\begin{aligned} commonIC(c_1, ..., c_n) &= IC(\bigcap_{i=1}^{n} c_i) \\ &\approx \max_{c \in subsum(c_1, ..., c_n)} [-\log P(c)] \end{aligned}$$

   where $subsum(c_1, ..., c_n)$ is the set of concepts that subsume all the concepts of $c_1, ..., c_n$ in SIS.

Unsupervised Algorithm

1. To apply non-overlapping IC of sentences in STS evaluation, the semantic information space (SIS) is constructed, which employs the super-subordinate (is-a) relation from the hierarchical taxonomy of WordNet.

2. The space size of a concept is the information content of the concept.

3. SIS is not a traditional orthogonality multidimensional space, while it is the space with inclusion relation among concepts.

4. Sentences in SIS are represented as a real physical space instead of a point in vector space.

5. The computational complexity of calculation on the basis of inclusion-exclusion principle is of $O(2^n)$. It can be brought down to $O(n)$ by the algorithm presented:

   - $Root(c_i)$ indicates the set of paths, each path is the node list from c_i to the root in the nominal hierarchical taxonomy of WordNet. $Root(n)$ is the short form of $Root(c_1, ..., c_n)$.
   - $Set(p)$ is the set of nodes in path p.
   - $Root(n) = \{p_k | \forall p_k \in Root(c_i), \nexists p_t \in Root(c_j),\ Set(p_k) \subseteq Set(p_t).\ i = 1, 2, ..., n;\ j = 1, 2, ..., n\}$.
   - $|Root(c_i)|$ means the number of paths in $Root(c_i)$.
   - $HSN(c_i)$ expresses the set of nodes in any of path in $Root(c_i)$. $HSN(n)$ is the short form of $HSN(c_1, ..., c_n)$.

     Formally, $HSN(n) = \{c_k | c_k \in HSN(c_i).\ i = 1, 2, ..., n\}$.
   - $Depth(c)$ is the max depth from concept c to the root.
   - $Intersect(n{+}1|n) = \{c_i | \forall c_i \in \{Set(p_t) \wedge HSN(n)\},\ \nexists c_j \in \{Set(p_t) \wedge HSN(n)\},\ depth(c_i) \leq depth(c_j),\ p_t \in Root(c_{n+1});\ t = 1, ..., |Root(c_{n+1}|\}$.
   - $totalIC(c_1, ..., c_n)$ is the quantity of total information of n-concepts.
   - Information content gain from concept c_i is defined as:

$$ICG(c_i) = IC(c_i) - totalIC(Intersect(i|i-1))$$

6. Finally, the algorithm is given as:

Algorithm 1 $getTotalIC(S)$

Input : $S : c_i|\ i = 1, 2, ..., n;\ n = |S|$
Output : tIC: Total IC of input S

if $S = \Phi$ **then**
  return 0
end if
Initialize: $tIC \leftarrow 0, Root(0) \leftarrow \Phi$
for $i = 1; i \leq n; i{+}{+}$ **do**
  $Intersect(i|i-1), Root(i) \leftarrow getIntersect(c_i, Root(i-1))$
  $ICG \leftarrow IC(c_i) - getTotalIC(Intersect(i|i-1))$
  $tIC \mathrel{+}= ICG$
end for
return tIC

7. It can be seen that the computational complexity of Algorithm 1 is $O(n)$.

Improving Recall Rate

1. WordNet is utilized to directly obtain the nominal forms of a content word which is not a noun mainly through derivational pointers in WordNet.

   The word formation helps enhance the recall rate of known content words in sentence-to- SIS mappings.

Algorithm 2 $getIntersect(c_i, Root(i-1))$

Input : $c_i, Root(i-1)$
Output : $Intersect(i|i-1), Root(i)$
Initialize: get $Root(c_i)$ from WordNet, $Intersect(i|i-1) \leftarrow \Phi$, $Root(i) \leftarrow Root(i-1)$
if $Root(i) = \Phi$ **then**
    $Root(i) \leftarrow Root(c_i)$
    return $Intersect(i|i-1), Root(i)$
end if
for each $r_i \in Root(c_i)$ **do**
    $pos \leftarrow depth(r_i) - 1$ ▷ pos ⇔ root
    for each $r_{i-1} \in Root(i-1)$ **do**
        $(p, q) \leftarrow$ deepest common node
        position: p in r_i, q in r_{i-1}
        if $p = 0$ **then** ▷ $r_i in r_{i-1}$
            add c_i to $Intersect(i|i-1)$
            break the outer foreach loop
        end if
        if $q = 0$ **then** ▷ $r_{i-1} in r_i$
            remove r_{i-1} from $Root(i)$
        end if
        if $p < pos$ **then** ▷ r_{i-1} intersect at deeper node in r_i
            $pos \leftarrow p$
        end if
    end for
    add r_i to $Root(i)$
    add $c_{pos} \in r_i$ to $Intersect(i|i-1)$
end for
return $Intersect(i|i-1), Root(i)$

2. Name entity (NE) recognition tool and the alignment tool are employed to obtain non-overlapping unknown NEs, which are used for simulating non-overlapping IC in SIS.
3. Finally, sentence IC is augmented by word weights which could deem as the importance of words.

Experiments

1. SemEval 2017 STS task assesses the ability of systems to determine the degree of semantic similarity between monolingual and cross-lingual sentences in Arabic, English, Spanish and a surprise language of Turkish.
2. The shared task is organized into a set of secondary sub-tracks and a single combined primary track.
3. Each secondary subtrack involves providing STS scores for monolingual sentence pairs in a particular language or for cross-lingual sentence pairs from the combination of two particular languages.
4. The SemEval 2017 STS shared task contains 1750 pairs with gold standard (GS) out of total 2000 pairs from 7 different tracks.
5. Systems were required to annotate all the pairs and performance was evaluated on all pairs or a subset with GS in the datasets.
6. The GS for each pair ranges from 0 to 5.
7. The evaluation metric is the Pearson product-moment correlation coefficient (PCC) between semantic similarity scores of machine assigned and human judgements. PCC is used for each individual test set, and the primary evaluation is measured by weighted mean of PCC on all datasets.

Results

	Primary	Track 1	Track 2	Track 3	Track 4a	Track 4b	Track 5	Track 6
Run 1	0.6703	0.7535	**0.7007**	0.8323	0.7813	0.0758	0.8161	**0.7327**
Run 2	0.6662	**0.7543**	0.6953	0.8289	0.7761	0.0584	0.8222	0.7280
Run 3	**0.6789**	0.7417	0.6965	**0.8499**	**0.7828**	**0.1107**	**0.8400**	0.7305
Cosine Baseline	0.5370	0.6045	0.5155	0.7117	0.6220	0.0320	0.7278	0.5456
Best System	0.7316	0.7440	0.7493	0.8559	0.8131	0.3363	0.8518	0.7706
All Single Best	-	0.7543	0.7493	0.8559	0.8302	0.3407	0.8547	0.7706
Differences	5.3%	-0.8%	4.9%	0.6%	4.7%	23.0%	1.5%	3.8%
Team Rankings	2	1	2	2	3	14	4	2

Table 2: Performances on SemEval 2017 STS evaluation datasets.

Set	Size	Run 1	Run 2	Run 3
Development	1500	0.8194	0.8240	**0.8291**
Test	1379	0.7942	0.7962	**0.8085**

Table 3: Performances of runs on STS benchmark.

Bibliography

1. 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). Association for Computational Linguistics, pages 252–263. https://doi.org/10.18653/v1/S152045.

2. 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). Association for Computational Linguistics, pages 81–91. https://doi.org/10.3115/v1/S14-2010.

3. Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1–14. http://www.aclweb.org/anthology/S17-2001.

4. Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, 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). Association for Computational Linguistics, pages 497–511. https://doi.org/10.18653/v1/S16-1081.

5. Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):27.

6. Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *SEM 2012: 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 2012). Association for Computational Linguistics, pages 385–393. http://aclweb.org/anthology/S12-1051.

7. Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UM-BC EBIQUITY-CORE: Semantic Textual Similar- ity Systems. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Association for Computational Linguistics, pages 44–52. http://aclweb.org/anthology/S13-1005.

8. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1):10–18.

9. Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD) 2(2):10.

10. 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: Semantic Textual Similarity. Association for Computational Linguistics, pages 32–43. http://aclweb.org/anthology/S13-1004.

11. Paul Jaccard. 1908. distribution florale. Nouvelles recherches sur la Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference Research on Computational Linguistics (ROCLING X).

12. Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *SEM 2012: 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 2012). Association for Computational Linguistics, pages 435–440. http://aclweb.org/anthology/S12-1059.

13. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, pages 177–180. http://aclweb.org/anthology/P07-2045.

14. Steven Bird. 2006. NLTK: The Natural Language Toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. Association for Computational Linguistics, pages 69–72. http://aclweb.org/anthology/P06-4018.

15. Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering 18(8):1138–1150.

16. Tomáš Brychcin and Lukáš Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic Textual Similarity using Lexical, Syntactic, and Semantic Information. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, pages 588–594. https://doi.org/10.18653/v1/S16-1089.

17. Dekang Lin. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In 35th Annual Meeting of the Association for Computational Linguistics. http://aclweb.org/anthology/P97-1009.

18. Dekang Lin. 1998. An information-theoretic definition of similarity. In ICML. Citeseer, volume 98, pages 296–304.

19. Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, pages 55–60. https://doi.org/10.3115/v1/P14-5010.

20. Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In AAAI. volume 6, pages 775–780.

21. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

22. Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition ldc2011t07, 2011. URL https://catalog.ldc.upenn. edu/LDC2011T07.

23. Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. International Joint Conference on Artificial Intelligence (IJCAI).

24. Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung poland nlp team at semeval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, pages 602–608. https://doi.org/10.18653/v1/S16-1091.

25. Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *SEM 2012: 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 2012). Association for Computational Linguistics, pages 441–448. http://aclweb.org/anthology/S12-1060.

26. Shirish Krishnaj Shevade, S Sathiya Keerthi, Chiranjib Bhattacharyya, and Karaturi Radha Krishna Murthy. 2000. Improvements to the smo algorithm for svm regression. IEEE transactions on neural networks 11(5):1188–1193.

27. Arafat Md 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). Association for Computational Linguistics, pages 241–246. https://doi.org/10.3115/v1/S142039.

28. Arafat Md Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. Transactions of the Association of Computational Linguistics 2:219–230. http://aclweb.org/anthology/Q14-1018.

29. Arafat Md 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 (SemEval 2015). Association for Computational Linguistics, pages 148–153. https://doi.org/10.18653/v1/S15-2027.

30. Hao Wu and Heyan Huang. 2016. Sentence similarity computational model based on information content. IEICE TRANSACTIONS on Information and Systems 99(6):1645–1652.

31. Hao Wu and Heyan Huang. 2017. Efficient algorithm for sentence information content computing in semantic hierarchical network. IEICE TRANSACTIONS on Information and Systems 100(1):238–241.