A Hybrid Approach for Sentence Similarity: Combining Semantic and Structural Similarity Metrics

Intelligent Systems, Delft University of Technology

Author: W.G. Haakman (w.g.haakman@student.tudelft.nl)
Supervisor: P.K. Murukannaiah (P.K.Murukannaiah@tudelft.nl)

TU Delft Delft University of Technology

GitLab Repository (request access)

1. Human Psychology

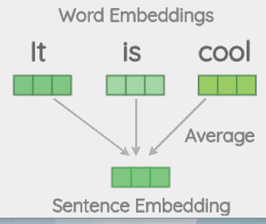
Sentences are composed of relations.
Relational similarity implies sentence similarity [1].

Do you want to come with us to the pub behind the hill?	Human Rating of Sentence Similarity: **0.455**
We are going out for drinks tonight in Salford Quays if you would like to come.	

2. Word-Based Approaches

Using Word-Embeddings like Word2Vec or GloVe.

Cosine Similarity of average feature vectors. [2]



3. Structure-Based Approaches

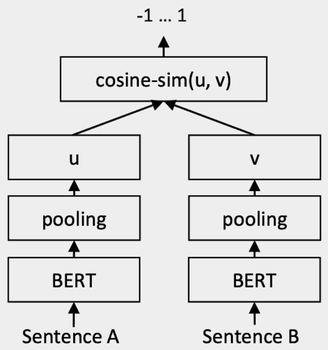
Take advantage of structure of a sentence; e.g. Word Order Distance [3].

$$r_1 = \{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\}$$
$$r_2 = \{1\ 2\ 3\ 9\ 5\ 6\ 7\ 8\ 4\}$$

$$S_r = 1 - \frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\|\mathbf{r}_1 + \mathbf{r}_2\|}$$

4. Vector-Based Approaches

Represent a sentence as a vector. Then calculate cosine similarity between sentence vectors;
e.g. SentenceBERT [4].



5. A Novel Approach

Smooth Inverse Frequency Semantic Similarity:

$$v_s(s) = \frac{1}{n}\sum_{w \in s}^{n} \frac{\alpha}{\alpha + zipf(w)} \cdot w2v(w)$$

$$sim_{sem}(s_1, s_2) = \frac{v_s(s_1) \cdot v_s(s_2)}{||v_s(s_1)|| \cdot ||v_s(s_2)||}$$

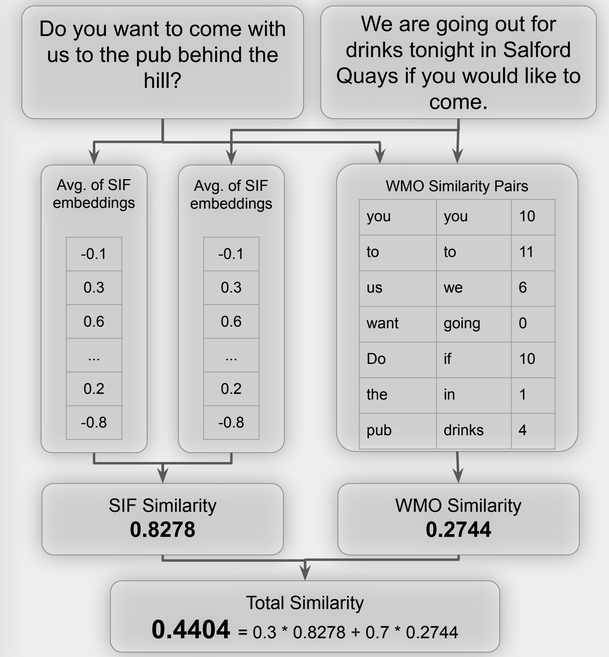
Word Mover's Order Similarity:

$$sim_{wmo}(s_1, s_2) = \frac{1}{n} \sum_{p_{min}(w_1, w_2)}^{n} \frac{1}{1 + \kappa \cdot dist_{wo}(w_1, w_2)} \cdot w2v(w_1, w_2)$$

Total Similarity:

$$sim(s_1, s_2) = \iota \cdot sim_{sem}(s_1, s_2) + (1 - \iota) \cdot sim_{wmo}(s_1, s_2)$$

$\alpha = 0.6876$, $\kappa = 0.09611$ and $\iota = 0.3$



6. Results STSS-131

Evaluation of proposed algorithm against SentenceBERT, LSA, and STASIS on STSS-131 [5].

Sentence Similarity Measures	STASIS	LSA	Sentence-BERT	**Proposed Algorithm**
Pearson Correlation Coefficient	0.636	0.693	0.936	**0.721**

7. Results MSRP

Evaluation of proposed algorithm against SentenceBERT and other common approaches on MSRP [6].

Sentence Similarity Measures	Prec.	Rec.	F_1	Acc.
$sim_{jaccard}$	0.835	0.603	0.7	0.657
sim_{sem}	0.674	0.99	0.802	0.675
sim_{wo}	0.681	0.619	0.648	0.554
sim_{sem+wo}	0.674	0.977	0.8	0.671
SentenceBERT	0.687	0.994	0.812	0.694
Proposed Algorithm	**0.742**	**0.856**	**0.795**	**0.715**

[1] https://doi.org/10.1016/j.cogpsych.2006.09.004
[2] https://openreview.net/pdf?id=SyK00v5xx
[3] https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-139.pdf
[4] https://www.aclweb.org/anthology/N19-1423.pdf
[5] https://doi.org/10.1145/2537046
[6] https://doi.org/10.1007/978-3-540-85836-2_29