A Statistical and Semantic Approach to Sentence Similarity

Anis Zaman **Advisor:** Rebecca Thomas
Department of Computer Science, Bard College, Spring 2013

Goal

- Explore different approaches for computing English sentence similarity
- Construct a large corpus of language Knowledge database

Motivation

- 56.6% of the data in the World Wide Web are in English
- 27% of Internet users browse in English
- Need robust Information Retrieval (IR) systems

Background

- **Data source**
  - Simple English Wikipedia and traditional Wikipedia articles
- **Term frequency and inverse document frequency (*tf-idf*)**
  - Idea: Less frequent words are more informative than commonly occurring ones
  - *Term frequency:* number of times the term appears in the document

$$tf(t,d) = \frac{frequency(t,d)}{\max\{frequency(w,d) : w \in d\}}$$

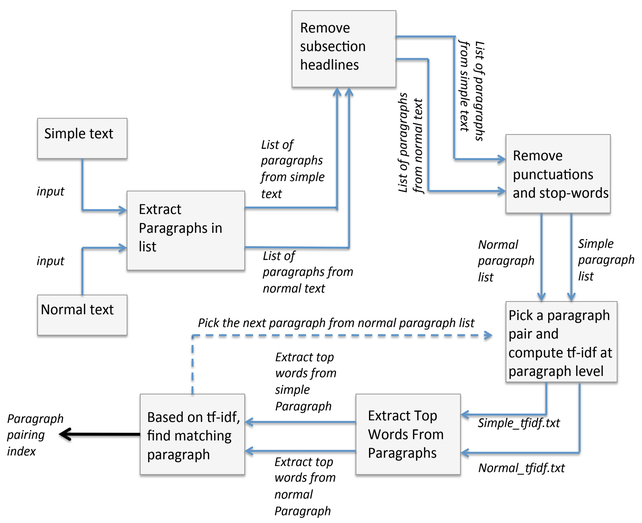
  - *Inverse document frequency:* measure of how rare or common a particular term is in a collection of documents.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

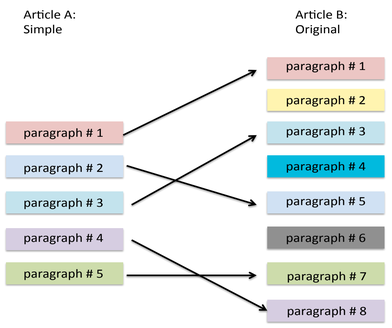
$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D)$$

Paragraph Alignment

Flow diagram for Paragraph Alignment



Example of Paragraph alignment

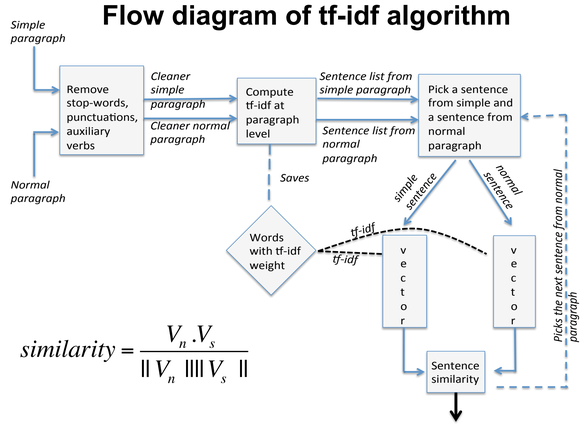


Sentence Similarity

Pure *tf-idf*

- Purely statistical

Flow diagram of tf-idf algorithm



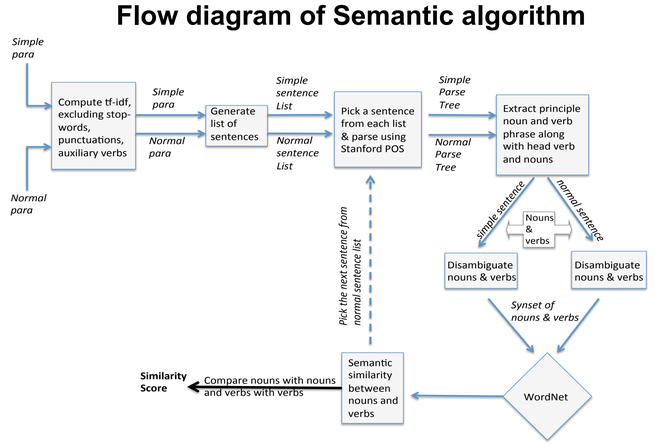
$$similarity = \frac{V_n . V_s}{\| V_n \| \| V_s \|}$$

Semantic

- Incorporates semantic distance between words from same POS

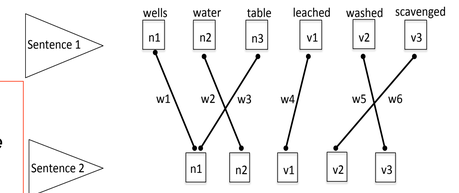
$$similarity(S_k, S_l) = \sum_{i=1}^{N} \max Sim(n_i \leftrightarrow n_j) + \sum_{i=1}^{V} \max Sim(v_i \leftrightarrow v_j)$$

Flow diagram of Semantic algorithm



Example Sentence

The wells and water table had been polluted by chemical pesticides that leached into the earth and washed by rain into the creeks, where the stunned fish were scavenged by the ospreys.



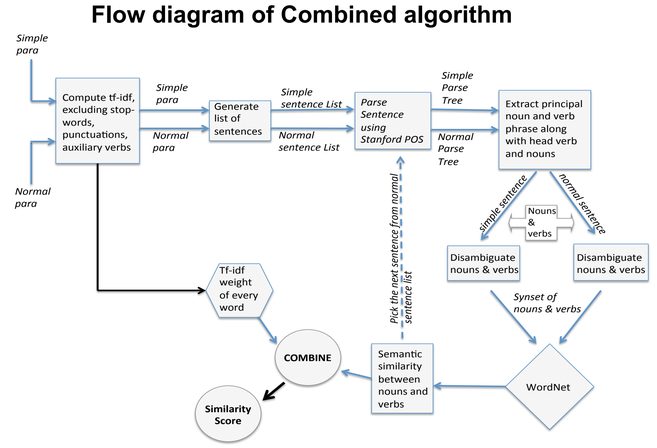
Semantic Similarity(Sentence1, Sentence 2) = w1 + w2 + w3 + w4 + w5 + w6

Combined

- Weights semantic distance with *tf-idf* weight

$$similarity(s,n) = \frac{1}{2}\left(\frac{\sum_{w \in s} \max Sim(w,n) \times tf - idf(w)}{\sum_{w \in s} tf - idf(w)} + \frac{\sum_{w \in n} \max Sim(w,s) \times tf - idf(w)}{\sum_{w \in n} tf - idf(w)}\right)$$

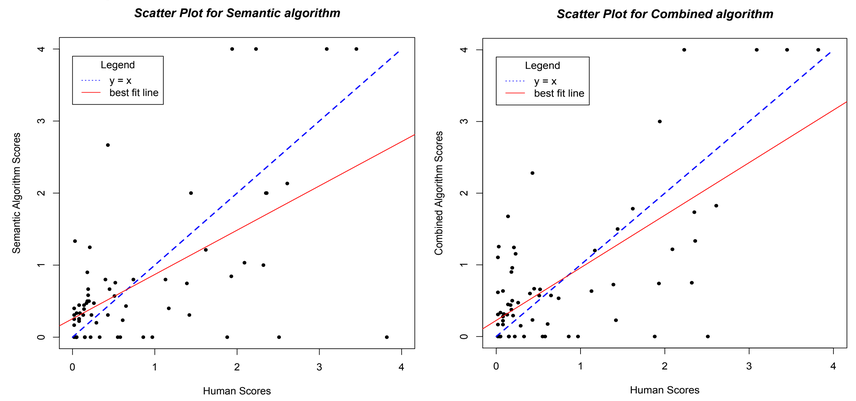
Flow diagram of Combined algorithm



Evaluation

Human judged sentences

- 65 human judged sentence pairs
- Scatter plot of Human score vs **Semantic** and **Combined** respectively

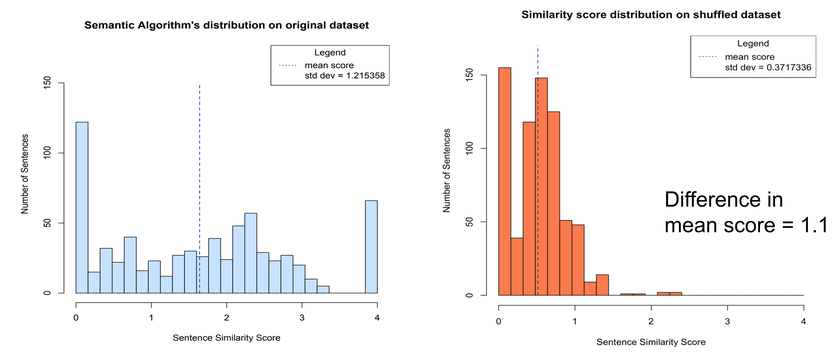


- Correlation for **Combined** is 0.68
- Correlation for **Semantic** is 0.58

SEMILAR paraphrase pairs

- 700 paraphrase sentence pairs, **Original**
- **Shuffled** dataset: Changed the second sentence of every pairs

Sentence similarity score distribution for both **Semantic** and **Combined** algorithm, respectively



Conclusion

- Similarity scores for **Combined** were closer to Human scores than **Semantic**
- Both good at detecting changes to the Original dataset
- **Combined** performs better in distinguishing between good aligned pairs from bad
- **Combined** is more promising than **Semantic**

Acknowledgement

I would like to thank Dr. Rebecca Thomas for advising me in this project

References

[Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633.
[Rus et al., 2012] Rus, V., Mihai Lintean, C. M., William Baggett, N. N., and Morgan, B. (2012). The similar corpus: A resource to foster the qualitative understanding of semantic similarity of texts. Language Resource Evaluation Conference.
[Coster and Kauchak, 2011] Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, pages 665–669.
[Mihalcea et al., 2006] Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the national conference on artificial intelligence, volume 21, page 775. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
[Li et al., 2006] Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. Knowledge and Data Engineering, IEEE Transactions on, 18(8): 1138–1150.