Detection of Intertextual References

Goal/Objective

- Identification of intertextual references, i.e. references made by one book to a list of potential candidate books.

  Example of biblical intertextuality:

- *For the scripture saith unto Pharaoh,* ***Even for this same purpose have I raised thee up, that I might shew my power in thee, and that my name might be declared throughout all the earth.*** [Romans 9:17; New Testament]
- ***And in very deed for this cause have I raised thee up, for to shew in thee my power; and that my name may be declared throughout all the earth.*** [Exodus 9:16; Old Testament]

Proposed Methodology

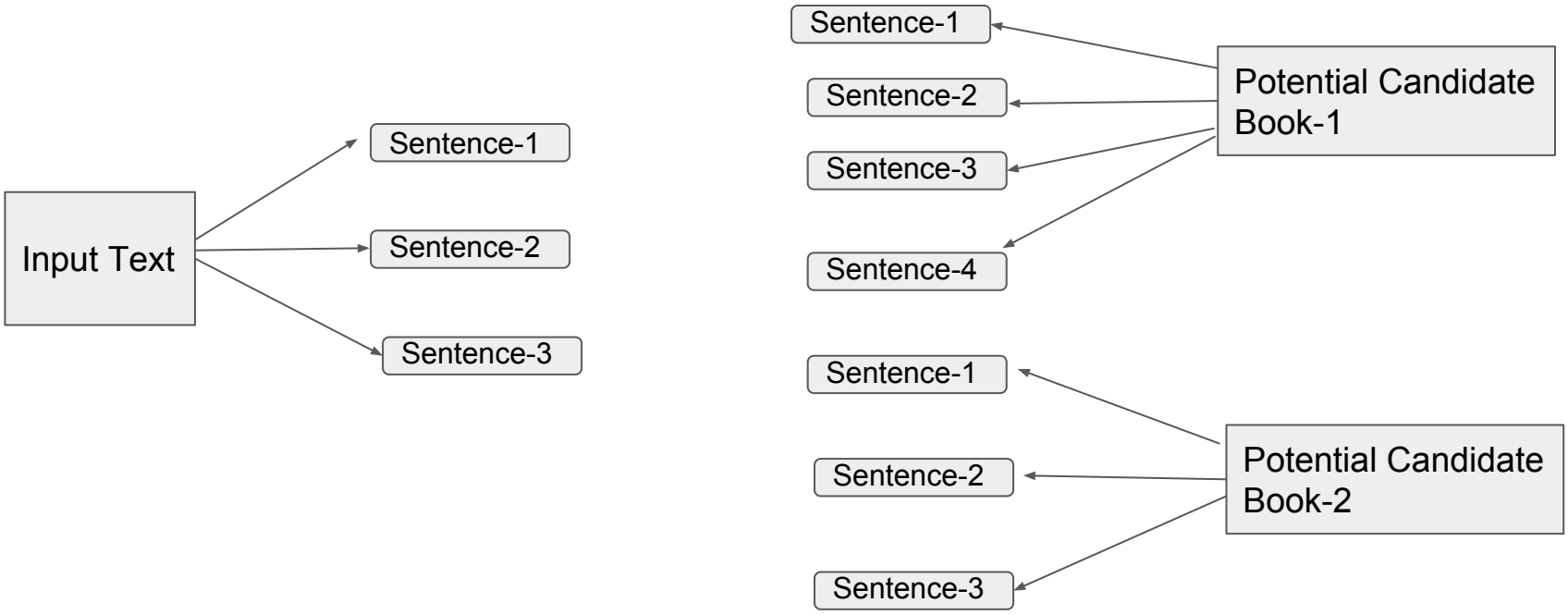
- New Text: The book which is making references to other texts. Example: A book written by Nietzsche.

- Potential Candidates: The list of books which the new text might be referring to. Example: Nietzsche’s personal library.

- Find the most similar sentence pairs between the new text and the potential candidates.

- Similarity is determined by both syntactic as well as semantic similarity.

Step 1: Splitting into sentences



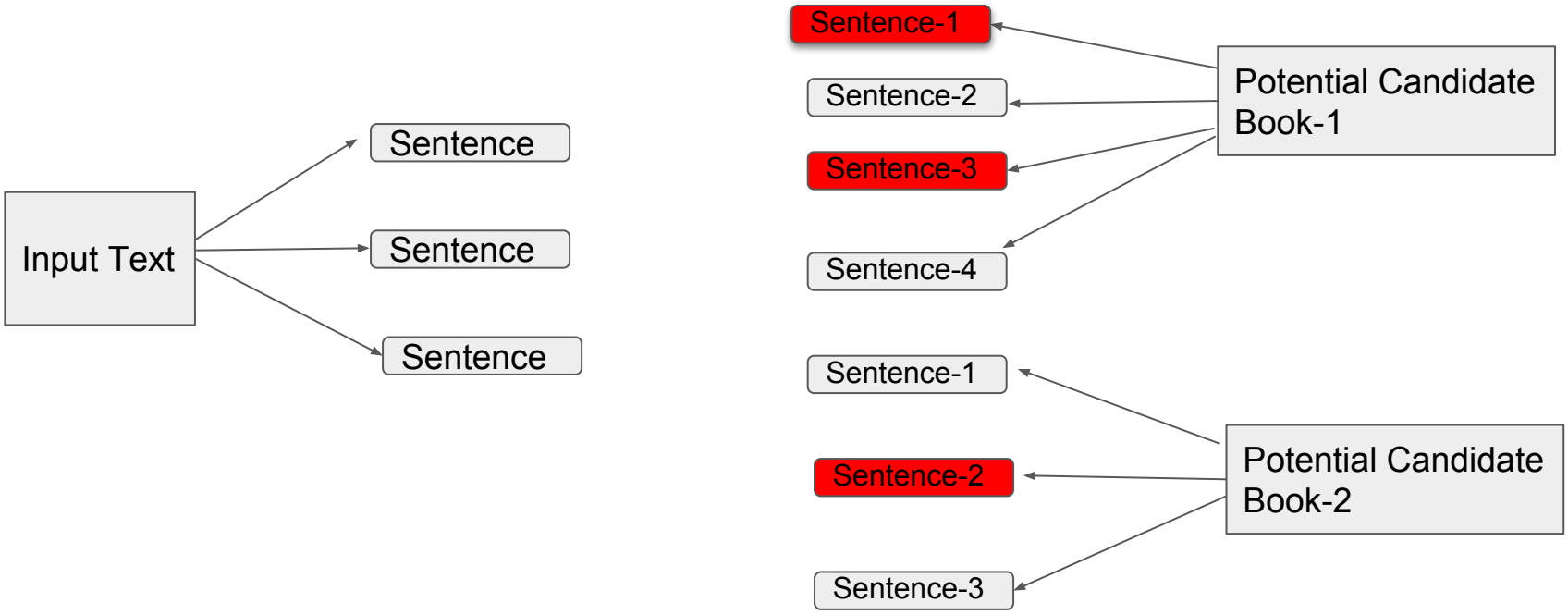
Step 2: Discarding Irrelevant Sentences

- Every sentence in the potential list is scored against every other sentence in the new text.

- The scoring is done using the jaccardian index, i.e. the number of common words in both the sentences.

- If a potential sentence is not similar (i.e. the jaccardian index is lower than a threshold - user defined parameter) to any of the sentences in the new text, then we discard the sentence

After discarding the irrelevant sentences



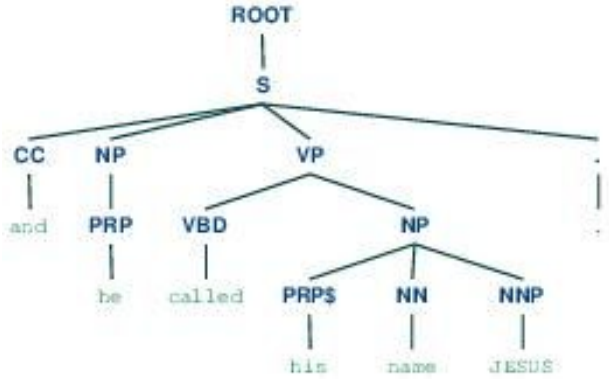
Step 3: Syntactic Similarity

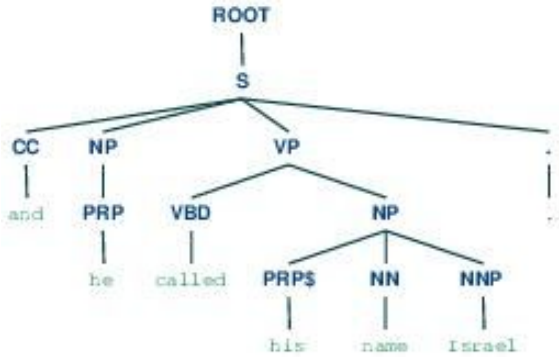
- Parse every sentence in the new text as well as all the remaining sentences in the potential candidates.

- Every sentence in the new text is scored against all the sentences in the potential candidates by comparing the similarity of their parse trees, i.e. of their syntactic structure.

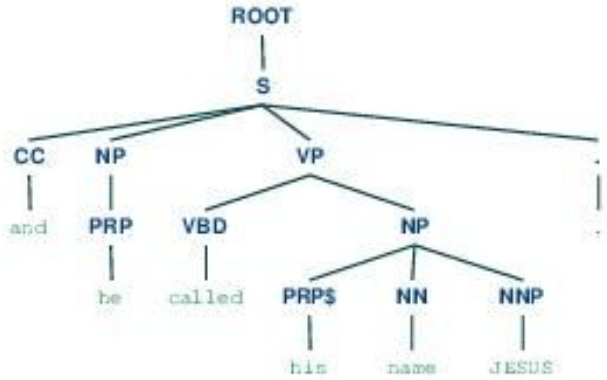
- We use a scoring technique called the Moschitti Score to compare their syntactic parse trees.

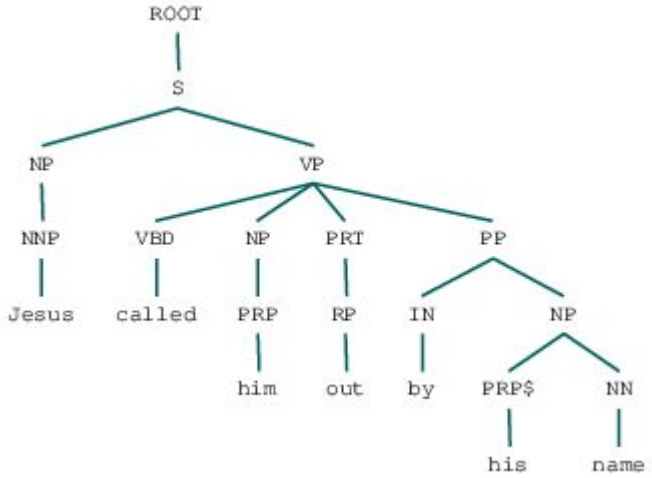
High Syntactic Similarity





Not very high syntactic similarity



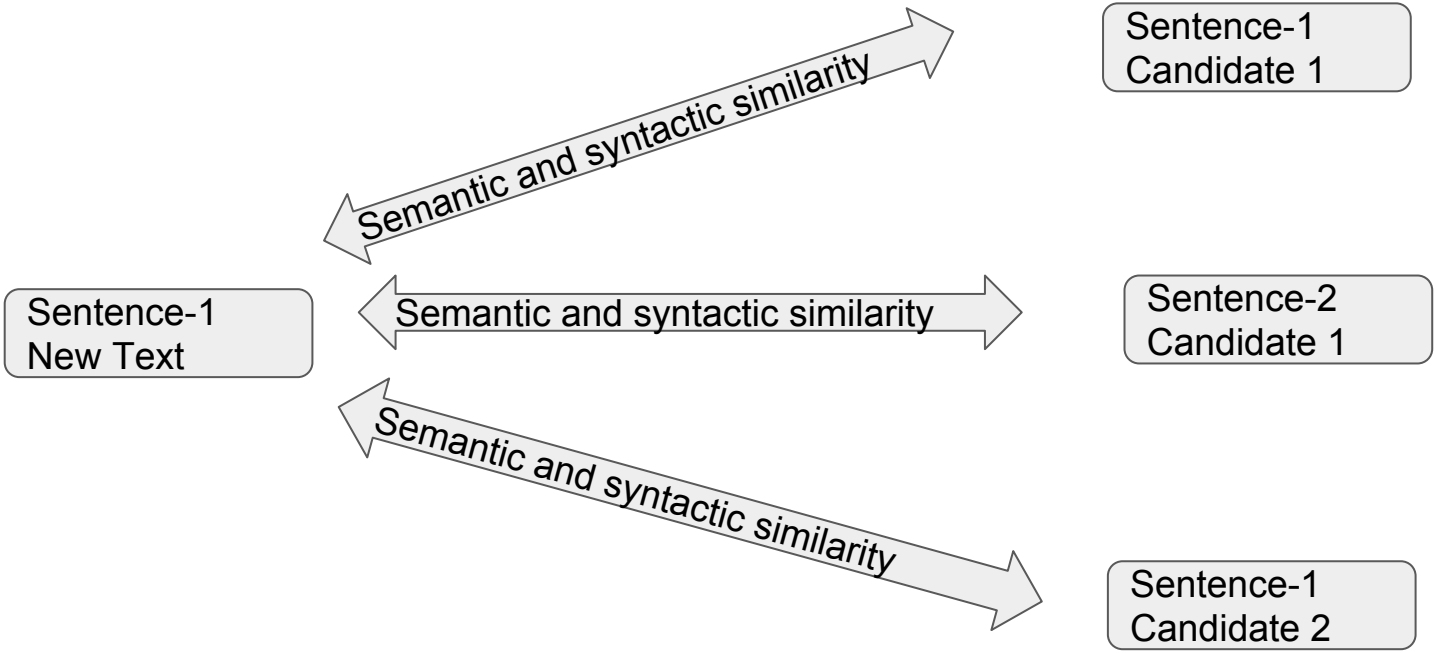


Step 4: Semantic Similarity

- Similarly, every sentence in the new text is scored against all the sentences in the potential candidates by comparing their semantic similarity.

- To compare their semantic similarity, we calculate the average word vector of both the sentences and determine the cosine similarity between these two word vectors.

Pairwise scoring of sentences



Step 5: Ranking the best pairs

- We return only those sentence pairs where the average syntactic and semantic similarity is greater than a threshold (another user-defined parameter).

- Finally, we rank the best pairs based on the number of overlapping nouns between the sentence pair.

Extension to Paragraph based intertextuality

- A similar system can be followed to compare paragraphs of the new text against paragraphs from the candidates.

- Semantic similarity is calculated using the average word vector of the entire paragraph.

- Syntactic similarity is calculated using the average syntactic similarity between sentence pairs.

Testing

- We tested our approach to search for biblical allusions made by the New Testament to sections of the Old Testament.

- New Text: New Testament

- Potential Candidates: Old Testament split into 29 books.

- The top 100 sentence pairs were presented to 4 annotators. Annotation task: Intertextual reference or not a reference

Testing

Annotators - Yes	Annotators - No	Number of Sentences
4	0	18
3	1	6
2	2	9

- These 33 sentence pairs ranked within the first 41 sentence pairs as per our final ranking.

Challenges / Issues

- Parameters that can be tuned:
  - Jaccardian threshold
  - Average syntactic and semantic similarity threshold

- Choices between similarity metrics:
  - Jaccardian index vs TF-IDF
  - Moschitti Score vs Other syntactic similarity metrics
  - Word2vec vs Other syntactic similarity metrics

- Final ranking based on Nouns?

Future Work / Improvements

- Concrete definition of intertextuality

- Data set to test the proposed methodology and tune the user defined parameters to optimum values.