Word Embeddings to Document Distances and Summaries

Vishwani Gupta

Supervisor: Sven Giesselbach

Goals :

1. Build and evaluate Word2Vec Model using different Text corpora.
2. Use the best model to implement Word Mover Distance.
3. Optimize the Word Mover Distance using three techniques proposed by Kusner et al.
    a. Word Centroid Distance
    b. Relaxed Word Mover Distance
    c. Prefetch and Prune
4. Use Word Mover Distance to explore different heuristics to generate summaries.

WORD2VEC

1. Helps in obtaining word vectors using context of the text.
2. Enable us to perform algebraic operations on the word vectors.
3. The words that are similar are closer to each other.
4. Using Google Model of Word2Vec, we obtain word vectors used in Word Mover Distance.

WORD MOVER DISTANCE AND OPTIMIZATION

1. WMD is used to obtain document distances.
2. Implicitly solves a linear optimization problem.

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

3. Three optimization approaches:
   a. Word Centroid Distance
   b. Prefetch and Prune
   c. Relaxed Word Mover Distance
4. WMD is used in extracting similar sentences which helps in generating summaries.

Summaries of documents

MOTIVATION:

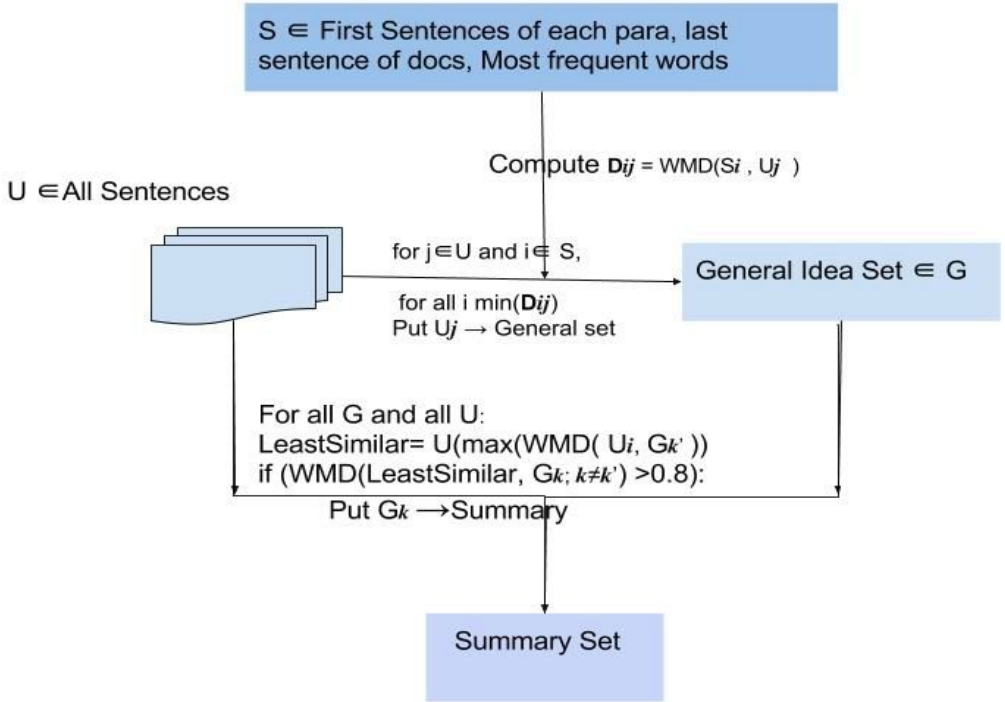
1. Gross et al. proposed an unsupervised approach, known as Association Mixture Text Summarization.
    a. Selects sentences from the document in a manner that they cover all the relevant information.
    b. The summarization task can be done in two steps:
        i. Computation of document specific associations.
        ii. Selection of sentences with strong word association.
2. Zhu et al. also proposed a technique in which one can choose the best sentences which not only cover the content but also avoid redundancy using relation among all the sentences.
    a. They used a rank function to measure the relevance of the sentence chosen.

References:
1. Document summarization based on word associations by Oskar Gross, Antonine Doucet, and Hannu Toivonen.
2. A novel relational learning to rank approach for topic-focused multi-document summarization. 2013 by Yadong Zhu, Yanyan Lan, Jiafeng Guo, Pan Du, and Xueqi Cheng.

OUR APPROACH

1. Unsupervised Approach.
2. According to Harvard Writing Center, the main idea of a document can be extracted using:
   a. First few sentences of every paragraph
   b. Last few sentences of the document
   c. Most frequent words
3. We extract sentences which are similar to these sentences from the whole document using Word Mover Distance.
4. To cover the details which are left out, we again use Word Mover Distance but compare the all sentences to the extracted sentences for dissimilarity.

Summary using General idea set which is generated from the whole text document



RESULTS

Original text length : 71 sentences

The sources are web link about Conservation of Sea Turtles:

1.http://wenku.baidu.com/view/23667425482fb4daa58d4b46.html

2.http://conserveturtles.org/turtleblog/blog/2016/07/26/tour-de-turtles-competitor-esperanza-spotted-nesting-in-mexico

3.http://www.conserveturtles.org/seaturtleinformation.php?page=conservation

Results:

Summary length: 6 sentences

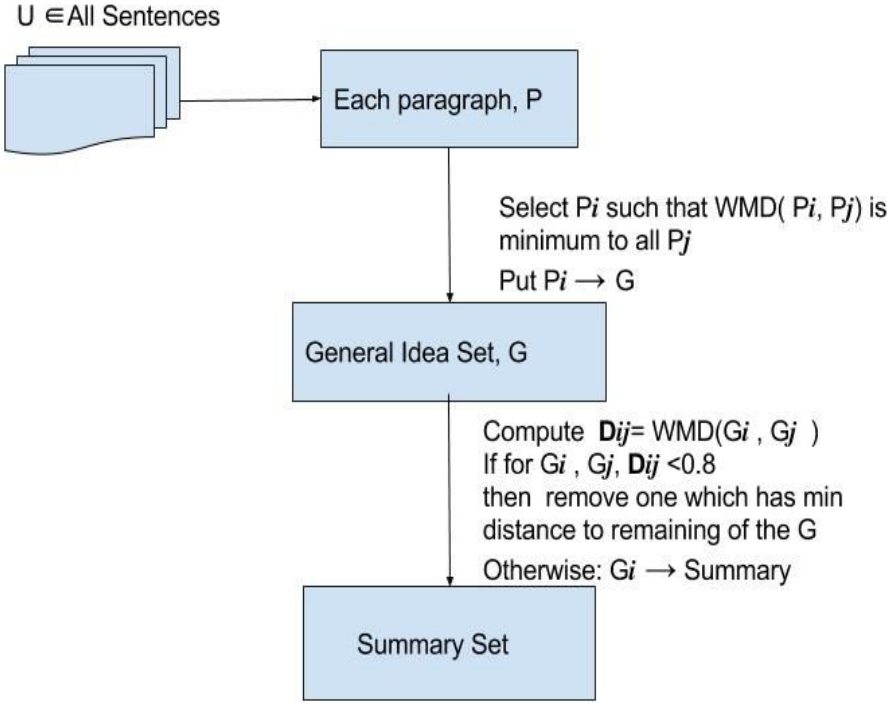
In many states where sea turtles nest, state laws have been passed to protect the species. The threats facing sea turtles are numerous and, for the most part, humans are the problem. For those of us trying to protect sea turtles, it is a mixed blessing that so many threats are human caused. The intrusion of tourists into these places make it difficult for the turtles to lay their eggs. The indiscriminate collection of turtles eggs on the beaches is no more allowed. One of the benefits of using satellite transmitters to track sea turtles is that it helps us determine turtles nesting site fidelity.

NOTE: The text is arranged manually for better understanding

SECOND APPROACH

1. We try to extract general idea of each paragraph of the documents.
2. Use Word Mover Distance to get most similar sentence in a paragraph.
3. Combine the sentences from each paragraph to get general idea set of the document.
4. Again use Word Mover Distance to remove redundancyfrom the document.

Summary using General idea set, generated from one paragraph at a time.



RESULTS

Original text length : 71 sentences

The source is again web sources about Conservation of Sea Turtles:

1.http://wenku.baidu.com/view/23667425482fb4daa58d4b46.html

2.http://conserveturtles.org/turtleblog/blog/2016/07/26/tour-de-turtles-competitor-esperanza-spotted-nesting-in-mexico

3.http://www.conserveturtles.org/seaturtleinformation.php?page=conservation

RESULTS

Original text length : 71 sentences

Summary length: 5 sentences

In many states where sea turtles nest, state laws have been passed to protect the species. For those of us trying to protect sea turtles, it is a mixed blessing that so many threats are human caused. The indiscriminate collection of turtles eggs on the beaches is no more allowed. The intrusion of tourists into these places make it difficult for the turtles to lay their eggs. Pollution of the sea has also reduced the number of turtles.

NOTE: The text is arranged manually for better understanding

EVALUATION

1. Both the approaches although have different sentences from the text, are similar in context.
2. One of the major drawback of these approaches is that we are picking the sentences from the text, it is not generated by the algorithm.
3. Another drawback is that the text is not in a flow, since it selects sentences which can be from any part of the document.

CONCLUSION

1. Different models of Word2Vec have been compared and it is seen that for getting a better model, more training data is always better.
2. In Word Mover Distance algorithm, we explore all the optimizations proposed by Matt Kusner et al. The optimization techniques gave very similar results to the Word Mover Distance and the time taken is almost one third of WMD.
3. Word Mover Distance is used to generate General Idea of the text.
4. Though the summaries generated from both approaches, need improvement as they lack flow of context and are composed of sentences from the text, we can nevertheless use Word Mover Distance and Word2Vec to get a general idea and then summarize the text.

Thank you