Document Summarization from WordNet Subgraph obtained by Recursive Gloss Overlap - Draft

$SrinivasanKannan(alias)Ka.Shrinivaasan(alias)ShrinivasKannan$
$IndependentOpenSourceDeveloper, ResearcherandConsultant$
Ph : 9789346927, 9003082186, 9791165980
$KrishnaiResearchOpenSourceProducts$: $http$: $//sourceforge.net/users/ka\_shrinivaasan$,
$https$: $//www.ohloh.net/accounts/ka\_shrinivaasan$
$ResearchWebsite$: $https$: $//sites.google.com/site/kuja27/$
$(ka.shrinivaasan@gmail.com, shrinivas.kannan@gmail.com, kashrinivaasan@live.com)$

July 25, 2014

Abstract

This article expands on the idea of construction of WordNet subgraph of a text document mentioned in references and studies the spectral properties of the WordNet subgraph and its Summarizability

1 Document Summarization and Recursive Gloss Overlap Algorithm

1. WordNet Subgraph is created for a document using Recursive Gloss Overlap algorithm as mentioned in the references below. Nodes of this graph are words and edges are relations between words. Breadth First Search or Depth First Search of this graph yields a flattened out meaning contained in the document. This just preserves the approximate meaning. The BFS or DFS prints "word1", "relation12" and "word2" iteratively for all nodes and edges. The grammar used for summary is just a concatenation of texts "word1", "relation12" and "word2".

2. Above BFS and DFS would find all the connected components of the WordNet subgraph. Thus summary of the document can be obtained by just concentrating on top few connected components (in terms of size). This is because size of a connected component in the WordNet subgraph implies the centrality or importance of that component to the document.

3. WordNet subgraph from Recursive Gloss Overlap is quite flexible as to what depth the definition graph or tree has to be grown. Once definition tree has been constructed, removal of isomorphic nodes or subtrees does away with duplication and gives a pruned wordnet subgraph with unique vertices and edges.

4. Mapping a text document to a graph as above is quite useful in applying lot of graph theorems and techniques on the wordnet subgraph.

5. As Document Summarization is basically an effort to highlight most important parts of a larger document, above graph precisely gives importance of a word to a document in terms of degree (incoming + outgoing edges) of that word vertex. More edges emanating from or connecting to a word implies that the word is a hub with edges as spokes to the document. This is an alternative measure to Term Frequency and Inverse Document Frequency for importance of a keyword. Moreover the degree is arrived after indepth growing of the Definition Tree which is tantamount to "human reasoning" or "studying in depth".Degrees of words could change if the depth to which the definition tree grown is altered which is commonsensical to human thinking process of in depth studying of a text yielding new insights.

6. Sorting the WordNet Subgraph nodes on the degree and sieving out nodes with less degrees is an effective summarization technique.

7. Random Walks on the WordNet subgraph of a document and their mixing time is a plausible measure of effort needed to understand a document (to get to Stationary Distribution)

8. Graph Spectra on the Laplacian of the WordNet Subgraph of a document could yield interesting details on the number of connected components (multiplicity of zero eigen value), spectral gap and algebraic connectivity (smallest and second smallest eigen values). Summarizability is inversely proportional to number of connected components and directly propertional to Algebraic Connectivity. If the WordNet Subgraph is more "Expanding" (i.e minimum of ratio of neighbours of all subset of vertices to the size of all subset of vertices) then that implies that the document has more internal connected structures and is summarizable.

2 Algorithm for Document Summarization from WordNet Subgraph

9. Algorithm for a summary (Theory only, without any experimental data):

   (a) Find word vertices of high degree above a threshold and their adjacent words and all paths amongst these high degree word vertex hubs. Add them to a new graph.

   (b) This gives a subgraph of high degree vertex hubs of WordNet Subgraph (Subgraph of a WordNet Subgraph).

   (c) Do BFS or DFS traversal and output sentences as "word(x)" "relation-xy" "word(y)" for all edges (x.y) iteratively for this high degree subgraph. This is the typical Hub-Spokes Model.

   (d) Without loss of generality, the spokes could have length of more than one edge. This captures the most important parts of a document and prunes others because high degree word vertices affect lot of other nodes in WordNet subgraph.

   (e) Hub words have more relations to adjacent words than non-Hub words in this graph.

10. Above algorithm can also optionally use PageRank algorithm to compute the ranks of highly important word vertices and choose only vertices above a threshold rank in constructing the subgraph of WordNet subgraph as above instead of degrees. (Importance of a word increases if it is pointed to by another important word). The graph is WordNet subgraph instead of hyperlink graph.

11. If the WordNet gives Parts-of-Speech information also, then more complex sentences can be constructed from the BFS or DFS traversal output. Probably this might need 2 passes - First pass creates simple sentences as above and creates a mapping table of the words to parts of speech during DFS or BFS and the second pass uses more complex sentence PoS template (e.g "Noun Pronoun Adjective Verb Obj") and looks up the table constructed in previous iteration to coalesce two or more simple sentences into complex sentences based on common substrings between two sentences. This coalescing can be hierarchically continued as a tree bottom-up to form more complex sentences from simple sentences.

12. Accuracy of the summary is dependent on the WordNet and also on the Summarizability measured by spectral properties above. If the WordNet is replaced by some other domain specific Ontology, relevance or meaningfulness of summary might increase. Above algorithm does not give much importance to very complex sentence formations. Only sentences of the form "word1 relates to word2" and their hierarchical coalitions are focussed. Complex sentences can be also obtained by transitive closure of the word-word relations. The Kleene closure of the above graph would give all possible such sentence formations. Above notwithstanding, the crucial meaning of the larger document is conveyed by the Summarization algorithm previously described.

3 Acknolwdgement

I dedicate this article to God.

4 Bibliography

References

[1] Algorithms for Intrinsic Merit of a Document - Recursive Gloss Overlap Algorithm for wordnet subgraph - $http://arxiv.org/abs/1006.4458$

[2] Slides iluustrating WordNet subgraph or Definition Graph Construction using Recursive Gloss Overlap Algorithm - $https://sites.google.com/site/kuja27/PresentationTAC2010.pdf?attredirects=0$

[3] Primitive implementation of the above at: $http://sourceforge.net/p/asfer/code/HEAD/tree/python-src/InterviewAlgorithm/$

[4] TAC 2010 - Update Summarization using Interview Algorithm - $http://www.nist.gov/tac/publications/2010/participant.papers/CMI_IIT.proceedings.pdf$

[5] WordNet - http://wordnet.princeton.edu/

[6] Other publication drafts on Majority Voting and Interview Algorithm related to the above in $https://sites.google.com/site/kuja27/$