Graph Based Methods For Text Summarization

Pallavi Gudipati
Gangal Varun Prashant

IIT Madras

March 20, 2015

Baselines

Centroid-based

- Baseline used in LexRank paper
- Non-graph based technique
- Centroid Score of a Word: The product of its tf in the whole document and its idf
- Centroid Score of a sentence: Summation of the centroid scores of the words contained

Global PageRank

Weighted version of PageRank algorithm on the whole sentence graph.

Global Influence Maximization

SimPath on the whole sentence graph with a fixed budget.

Approaches

- Community Detection followed by PageRank
- Community Detection followed by Influence Maximization

Implementation

- We use the classic CNM(Clauset-Newman-Moore) algorithm, based on greedy optimization of modularity, for community detection.
- Influence Maximization is performed under the Linear Threshold diffusion model, using the SimPath algorithm.
- The budgets for influence maximization/PageRank in each community are linearly proportional to the size of the community. However, the approach is extensible to other schemes(such as L1 or sqrt of size)

Evaluation

Datset

- TIPSTER's Computation and Language (cmp-lg) corpus
- 183 documents from cmp-lg collection and has been marked up in XML
- Scientific papers which appeared in Association for Computational Linguistics (ACL) sponsored conferences
- Extracted all the text present in its abstract and body. Abstract is used as a reference summary.
- Run our summarization algorithms on the text extracted from the body
- Considered only the papers that have abstracts with length more that 5 sentences. This leaves us with 20 papers in our dataset.

Evaluation

Evaluation Metric - ROUGE-1

- ROUGE - recall-based metric for fixed-length summaries which is based on n-gram co-occurrence.
- ROUGE-1 reports a score based on the 1-gram matching between the ground truth summaries and the summaries to be evaluated.

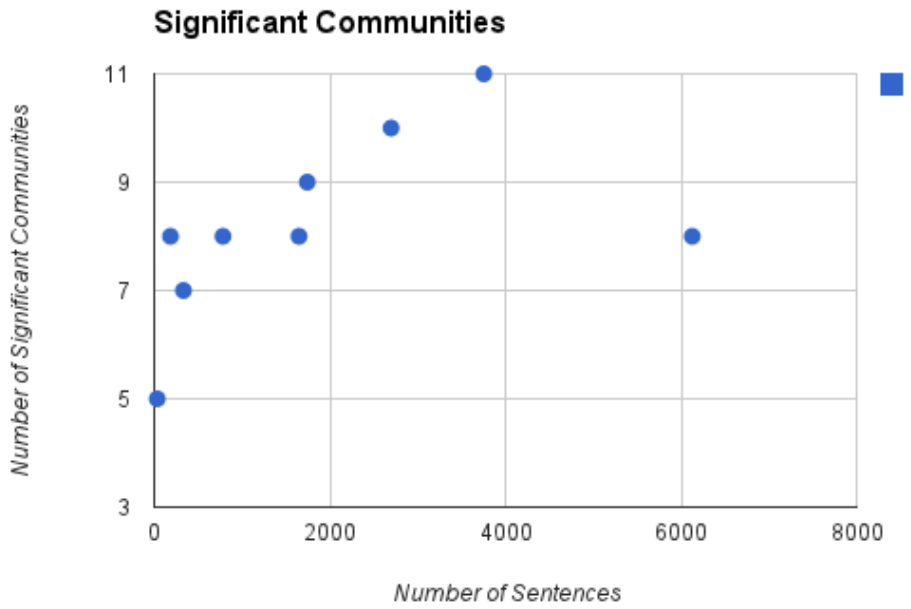
$$ROUGE - 1(s) = \frac{\sum_{r \in R} < \phi(r), \phi(s) >}{\sum_{r \in R} < \phi(r), \phi(r) >} \quad (1)$$

where R is the set of reference summaries and s is the summary that needs to be evaluates.

Observations

Significant Communities

The number of significant communities is always upper-bounded by a small constant. This shows that the number of latent concepts(or threads of perception) in a document that humans can perceive is upper-bound by a small number.



Observations

Performance against Baselines

Atleast one of our approaches outperforms all the baselines in every case(forward, backward or undirected). PageRank with CD gives the best overall performance.

Technique	**Avg. ROUGE-1 Score**		
	Undirected	*Forward*	*Backward*
Centroid-based	0.2460	0.2681	0.2595
Global PageRank	0.3520	0.3859	0.3722
Global Influence Maximization	0.1718	0.2342	0.2208
PageRank with CD	0.4307	0.3919	0.4335
Influence Maximization with CD	0.3312	0.4029	0.3779

Table: ROUGE-1 Scores

Observations

Directionality of Graph

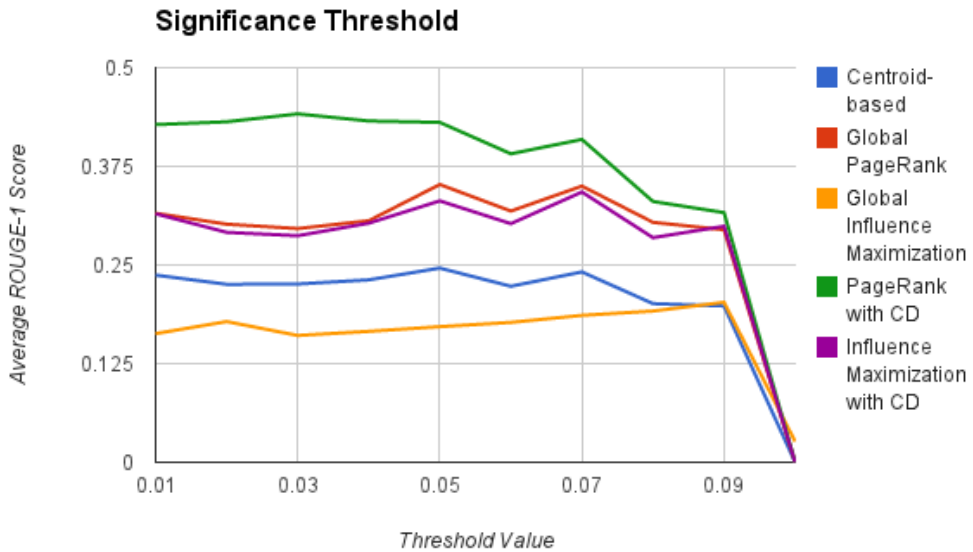
We observe that CD followed by IM best when the sentence graph is forward directed, and even performs better than CD followed by PageRank, and Global PageRank. This could be because

1. Directed graphs are more suited for influence maximization because the spread of influence in influence maximization is modelled in an inherently directed fashion.
2. Forward directed graphs suit influence maximization better than backward directed ones in the context of this dataset, since they accurately model the flow of information in the documents, each of which is a research paper. In a research paper represented by a forward directed graph, sentences in the initial sections (e.g Introduction) transmit their information to the latter sections(e.g Experiments).

Observations

Varying the Significance Threshold

Varying the significance threshold captures an important tradeoff between removing spurious edges and retaining significant ones. One needs to find the optimal value of the threshold through empirical tuning of the threshold. We find this threshold to be about 0.05 for most approaches, for the *TIPSTER* dataset.



Conclusions

- The approaches with Community Detection perform better than their global counter parts.
- Community Detection plays its role in discovering suppressed topics, and making sure that they have sufficient representation in the final summary.
- There are only a few significant communities to be discovered in a document's sentence graph, and the number of such communities remains fairly constant across the size of the document.
- Directionality of graph can have significant implications for the performance of influence maximization based methods, and has to be tailored to the requirements of the domain.
- In our case, we saw that forward directionality was favourable for the domain under consideration, i.e research papers.

Future Work

1. Using distributional similarity to model pairwise similarity between sentences. This will allow us to relate cases like the occurrence of the words **ship** in one sentence and **sea** in the other.
2. Experimenting with different schemes of budget allocation. For instance, we can adopt a regularized scheme of budget allocation(motivated by approaches such as L1 Regularized Logistic Regression in Machine Learning). This prevents very large communities from getting a disproportionate share of the summary.
3. Extending our approaches to the multi-document summarization setting, with both meta-summarization and complete graph summarization approaches to handle multiple documents.
4. Comparing the communities formed with concepts obtained through LSA.