HipoRank: Discourse-Aware Unsupervised Summarization Long Scientific Documents

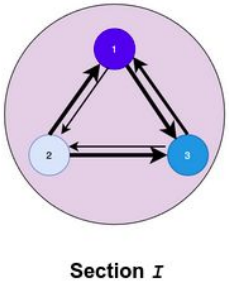
TL;DR

Motivation: leverage the discourse structure of scientific articles in unsupervised summarization.
Method: graph-based sentence ranking algorithm with two-level section hierarchy and directed edges weighted by asymmetric positional cues.
Results: performs much better than previous unsupervised approaches and comparable to many supervised model on PubMed/ArXiv datasets.
Takeaway: discourse structure is highly useful for determining sentence importance in scientific docs.

Incorporating Discourse Structure

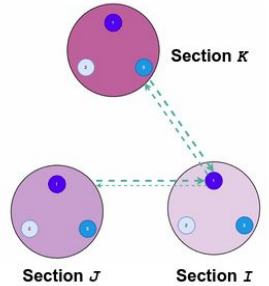
Intra-Section Edges

A sentence's importance depends on its relation to same-section sentences



Inter-Section Edges

A sentence's importance depends on its relation to other sections



Boundary Function

Edges are weighted more if they point to a sentence near a section start or end

$$c(v_i^I) = \mu_1 \cdot c_{\text{inter}}(v_i^I) + c_{\text{intra}}(v_i^I)$$

$$c_{\text{intra}}(v_i^I) = \sum_{v_j^I \in I} \frac{w_{ji}^I}{|I|} \quad c_{\text{inter}}(v_i^I) = \sum_{v^J \in D} \frac{w_i^{JI}}{|D|}$$

$$w_{ji}^I = \begin{cases} \lambda_1 * sim(v_j^I, v_i^I), & \text{if } d_b(v_i^I) \geq d_b(v_j^I), \\ \lambda_2 * sim(v_j^I, v_i^I), & \text{if } d_b(v_i^I) < d_b(v_j^I) \end{cases}$$

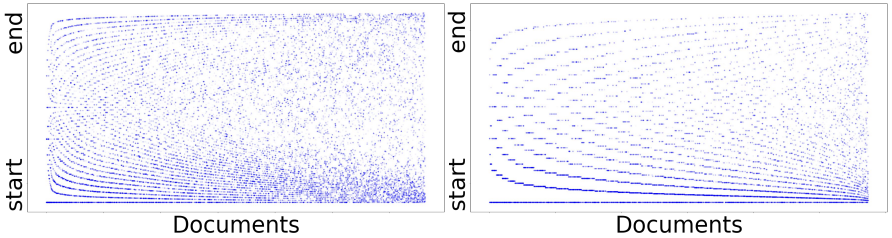
where $\lambda_1 < \lambda_2$

$$d_b(v_i^I) = \min(x_i^I, \alpha(n^I - x_i^I))$$

n^I is the number of sentences in section I
x_i^I represents sentence i's position in section I

Discourse Structure in Scientific Articles

Discourse structure suggested by sentence position (y-axis) and ROUGE-2 (color) is different in scientific articles (left, PubMed) and news articles (right, CNN)



Results on PubMed/ArXiv

Test set results on PubMed (ROUGE F1):

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead	35.63	12.28	25.17
Oracle (ROUGE-2, F1)	55.05	27.48	38.66
Unsupervised Extractive			
SumBasic (2007)	37.15	11.36	33.43
LSA (2004)	33.89	9.93	29.70
LexRank (2004)	39.19	13.89	34.59
PacSum (2019)	39.79	14.00	36.09
HipoRank (ours)	**43.58**	**17.00**	**39.31**

Test set results on arXiv (ROUGE F1):

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead	33.66	8.94	22.19
Oracle (ROUGE-2, F1)	53.88	23.05	34.90
Unsupervised Extractive			
SumBasic (2007)	29.47	6.95	26.30
LSA (2004)	29.91	7.42	25.67
LexRank (2004)	33.85	10.73	28.99
PacSum (2019)	38.57	10.93	34.33
HipoRank (ours)	**39.34**	**12.56**	**34.89**

Human Evaluation Results on PubMed

Model	Content-coverage	Importance
PacSum	30.52	48.70
HipoRank (ours)	**42.13**	**59.06**

Fleiss κ: 46.56/41.37 for content-coverage/importance respectively.
Coverage: Does it cover content from the abstract?
Importance: Is it important for a goal-oriented reader (Lin and Hovy,1997)?

Conclusions

Discourse structure is highly useful for determining sentence importance in scientific documents, which HipoRank leverages into a strong unsupervised baseline.