

ACL読み会

# **Sentence Centrality Revisited for Unsupervised Summarization**

Hao Zheng and Mirella Lapata  
University of Edinburgh

論文URL : <https://www.aclweb.org/anthology/P19-1628/>  
提案手法ソースコード : <https://github.com/mswellhao/PacSum>

紹介者：駒水 孝裕（名古屋大学）

# 概要

- やりたいこと：教師なし文書要約
- 提案手法：PacSUM  
(**P**osition-**A**ugmented **C**entrality based **S**ummarization)
  - Centrality-based Summarization
    - 文書を文をノードとする重み付きグラフで表現
      - 重み：文間類似度
    - グラフ上で重要な文を要約文として抽出
      - 重要度：次数中心性, PageRank, など
  - 文間類似度
    - 文表現にBERT
    - 類似度に文表現の内積 (cosine 類似度ではない)
  - 有向グラフ
    - 文の (相対的な) 出現位置を考慮：前に出てくる文ほど重要

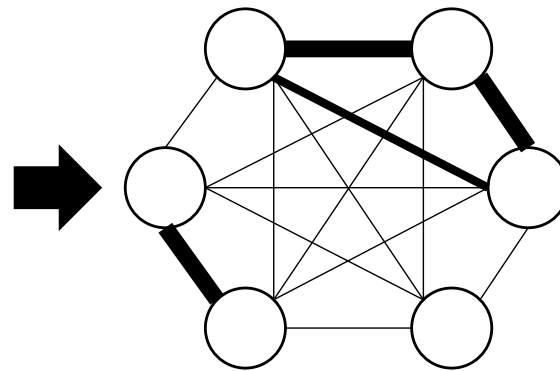
## 文書要約：教師あり v. 教師なし

教師あり	教師なし
<ul style="list-style-type: none"><li>• 性能・高</li><li>• 教師データが必要</li></ul>	<ul style="list-style-type: none"><li>• 性能・低</li><li>• 教師データが不要</li></ul>

- 近年の状況
  - 教師ありデータセット数が増加傾向
  - 言語や分野に網羅的なデータセットは望めない
- 教師なし文書要約は未だ重要
  - 性能向上が必須

# 教師なし文書要約：グラフアプローチ

- 代表的な手法：TextRank [Mihalcea+, 2004]



無向グラフ  
ノード：文  
エッジ：類似度

TF-IDFベクトルの  
cosine類似度

グラフの中心性に  
基づく文選択

- 次数中心性

$$centrality(s_i) = \sum_{j \neq i} e_{ij}$$

※  $e_{ij}$  は重み

- PageRank

文書

# PACSUM : 文間類似度

- BERT を fine-tuning

- 仮説「ある文の前後の文は関連が強い」
- 目的関数：

$$\log \sigma(v_{s_{i-1}}^T v_{s_i}) + \log \sigma(v_{s_{i+1}}^T v_{s_i}) + \mathbb{E}_s[\log \sigma(-v_s^T v_{s_i})]$$

前後の文とは近いベクトルに

他の文とは遠いベクトルに

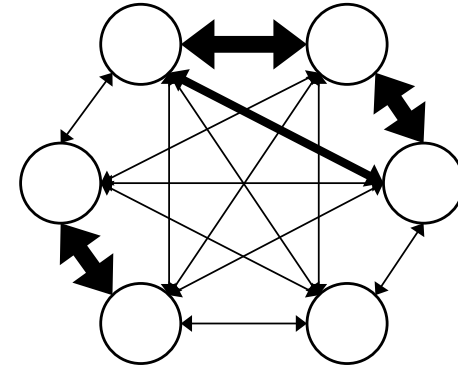
- 文間類似度

- 類似度： $\bar{E}_{ij} = v_i^T v_j$
- 正規化： $\tilde{E}_{ij} = \bar{E}_{ij} - [\min \bar{E} + \beta(\max \bar{E} - \min \bar{E})]$
- 枝刈り： $E_{ij} = \begin{cases} \tilde{E}_{ij} & \text{if } \tilde{E}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$ 
  - $\beta \in [0,1]$  が枝刈りのパラメタ

高すぎる値を補正  
(正規化ではない)

# PACSUM：有向グラフにおける中心性

- 有向性：文の相対的な前後関係
  - 無向エッジ → 双方向エッジ
- 入次数と出次数を区別した中心性

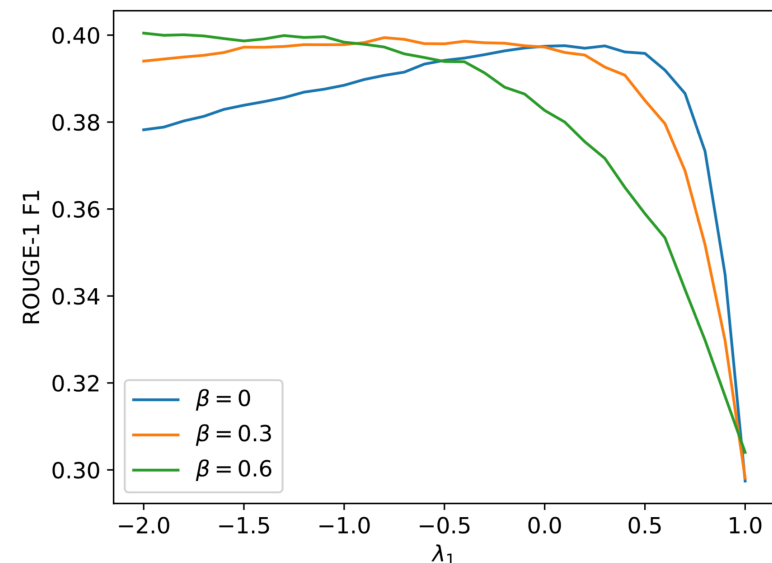


$$centrality(s_i) = \lambda_1 \sum_{j < i} e_{ij} + \lambda_2 \sum_{i < j} e_{ij}$$

前方の文（入次数）

後方の文（出次数）

- $\lambda_1 + \lambda_2 = 1$ 
  - $\lambda_1, \lambda_2$ ：重要度パラメタ
  - $\lambda_1$ は負値になりがち（右図）
    - 前の文との類似度は文書要約にネガティブ効果



# 実験：教師ありモデルに匹敵

※ ORACLE：最良のROUGE-?をなす文選択要約

		New York Times			CNN/DailyMail		
Method		NYT			CNN+DM		
		R-1	R-2	R-L	R-1	R-2	R-L
教師あり	ORACLE	61.9	41.7	58.3	54.7	30.4	50.8
	REFRESH <sup>4</sup> (Narayan et al., 2018b)	41.3	22.0	37.8	41.3	18.4	37.5
	POINTER-GENERATOR (See et al., 2017)	42.7	22.1	38.0	39.5	17.3	36.4
教師なし	LEAD-3	35.5	17.2	32.0	40.5	17.7	36.7
	DEGREE (tf-idf)	33.2	13.1	29.0	33.0	11.7	29.5
	TEXTRANK (tf-idf)	33.2	13.1	29.0	33.2	11.8	29.6
	TEXTRANK (skip-thought vectors)	30.1	9.6	26.1	31.4	10.2	28.2
	TEXTRANK (BERT)	29.7	9.0	25.3	30.8	9.6	27.4
提案手法	PACSUM (tf-idf)	40.4	20.6	36.4	39.2	16.3	35.3
	PACSUM (skip-thought vectors)	38.3	18.8	34.5	38.6	16.1	34.9
	PACSUM (BERT)	41.4	21.7	37.5	40.7	17.8	36.9

評価指標：ROUGE (1, 2, L)

- 教師なし学習で PACSUM が最も高性能
- 文表現をBERTにしたことで性能向上

## 実験：中国語のニュース要約

Method	TTNews		
	R-1	R-2	R-L
ORACLE	45.6	31.4	41.7
POINTER-GENERATOR	42.7	27.5	36.2
LEAD	30.8	18.4	24.9
TEXTRANK (tf-idf)	25.6	13.1	19.7
PACSUM (BERT)	32.8	18.9	26.1

- 英語のデータセットより正解要約が abstractive
  - Pointer-Generator が非常に良い
    - Pointer-Generator : abstractive method
    - その他 : extractive method
- リード法/TextRankよりは良い



# 出力結果例 (NYT)

## GOLD

Marine Corps says that V-22 Osprey, hybrid aircraft with troubled past, will be sent to Iraq in September, where it will see combat for first time.

The Pentagon has placed so many restrictions on how it can be used in combat that the plane – which is able to drop troops into battle like helicopter and then speed away like airplane – could have difficulty fulfilling marines longstanding mission for it.

Limitations on v-22, which cost \$80 million apiece, mean it can not evade enemy fire with same maneuvers and sharp turns used by helicopter pilots.

## PacSum

The Marine Corps said yesterday that the V-22 Osprey, a hybrid aircraft with a troubled past, will be sent to Iraq this September, where it will see combat for the first time.

The Pentagon has placed so many restrictions on how it can be used in combat that the plane — which is able to drop troops into battle like a helicopter and then speed away from danger like an airplane — could have difficulty fulfilling the Marines' longstanding mission for it.

The limitations on the V-22, which cost \$80 million apiece, mean it cannot evade enemy fire with the same maneuvers and sharp turns used by helicopter pilots.

## LEAD-3

the Marine Corps said yesterday that the V-22 Osprey, a hybrid aircraft with a troubled past, will be sent to Iraq this September, where it will see combat for the first time.

But because of a checkered safety record in test flights, the v-22 will be kept on a short leash.

The Pentagon has placed so many restrictions on how it can be used in combat that the plane – which is able to drop troops into battle like a helicopter and then speed away from danger like an airplane – could have difficulty fulfilling the marines' longstanding mission for it.

細かい言い回し以外はほぼ同じ

提案手法は二文目を除外

# 人力評価：QA方式

Method	NYT	CNN+DM	TTNews
ORACLE	49.0*	53.9*	60.0*
REFRESH	42.5	34.2	—
LEAD	34.7*	26.0*	50.0*
PACSUM	44.4	31.1	56.0

- やり方

- 正解要約から質問を作成
  - 質問：最重要コンテンツを答える質問
- 要約を見て人が回答
  - 十分な情報を含む要約が有用
- 正答率で評価（100点満点）
  - 部分正解：半分のスコア

- 結果（右上図）

- そもそもORACLEが良くない ➡ 文選択の限界
- 提案手法
  - 他手法より良い ➡ 改善の余地あり
  - ORACLEには劣る

# まとめ

- 教師なし文書要約手法：PacSUM  
(**P**osition-**A**ugmented **C**entrality based **S**ummarization)
  - Centrality-based Summarization
  - BERTに基づく文間類似度
  - 有向グラフ：文の（相対的な）出現位置を考慮
- 実験
  - 教師ありモデルに匹敵する性能
  - 教師なしモデルでは最良
    - ORACLEには及ばない ➡ 改善の余地あり
- なお、この論文の要約は提案手法で作られていない