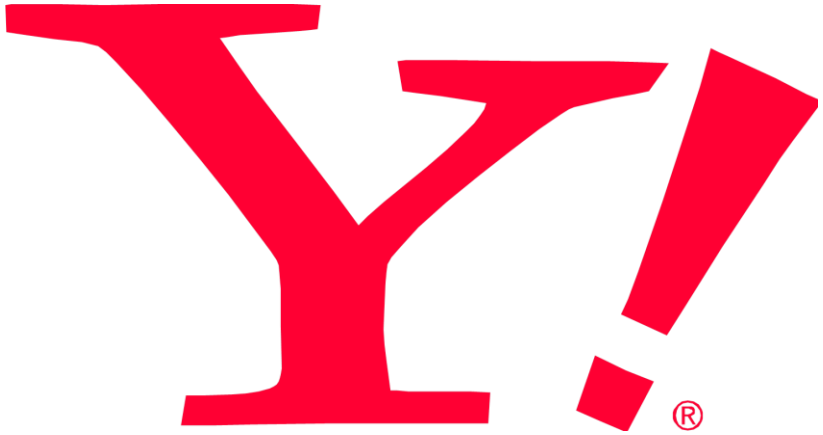


国際会議報告 EMNLP 2015
～文書要約～

小林隼人

2015/12/4

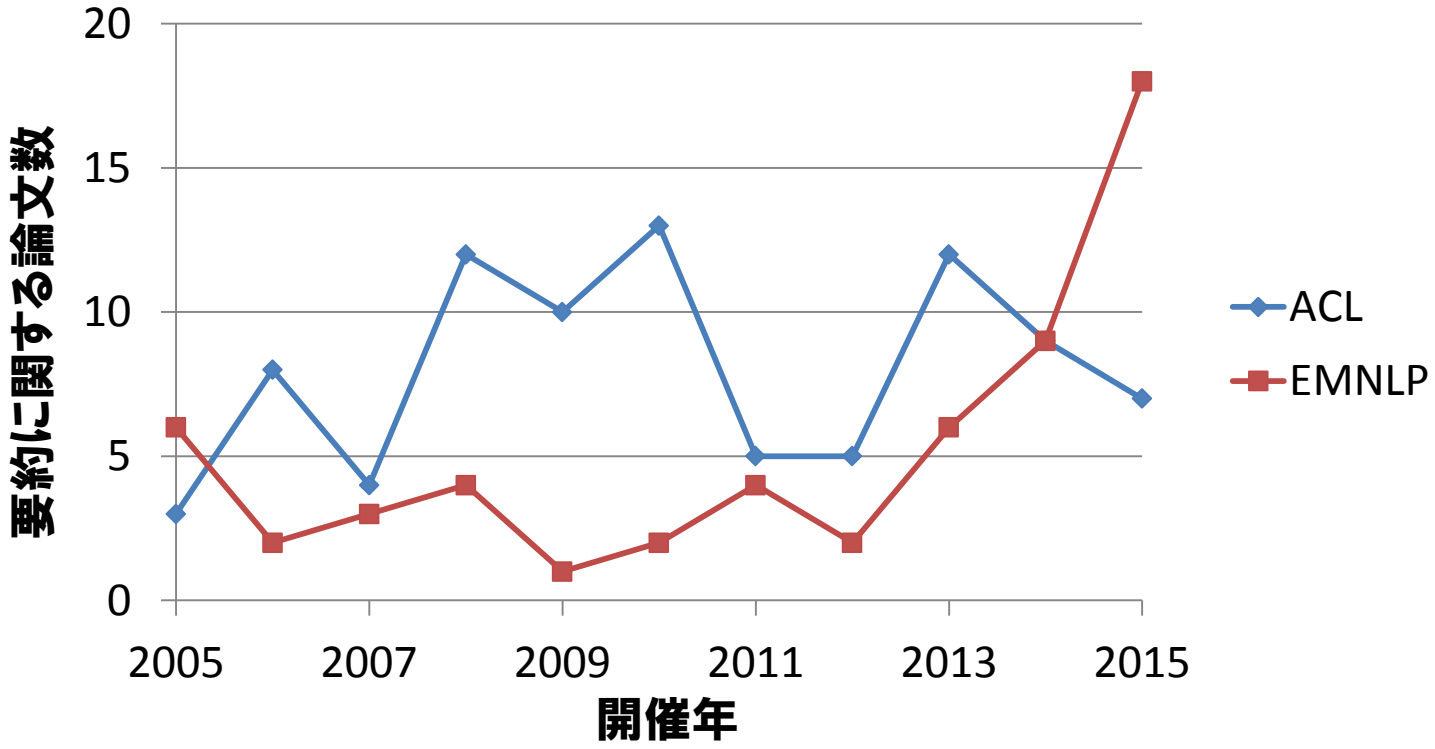
自己紹介

- 名前： 小林隼人(ハヤト・コバヤシ)
- 所属： Yahoo! JAPAN 研究所 ('13年入社)
  言語処理・機械学習室
- 略歴： 九大→東北大→東芝→ヤフー
- 研究歴： ロボット→学習理論→言語処理
- 最近の興味： 文書要約・生成
- 最近の成果(がんばってますアピール)
  - ACL'14, COLING'14, PACLING'15, WWW'15, ECML-PKDD'15, SIGDIAL'15, **EMNLP'15**, WSDM'16, …

今日は文書要約の研究を紹介します

- 要約研究の全体感
- 劣モジュラ最適化
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

- 文書要約に関する論文は１８本
  - 多いように見えますが、今年は発表数も2倍になっているので昨年と同程度？



全体感(2/3)

- 研究内容は大きく分けて3つ
  - 評価尺度2本、文生成3本、文抽出13本

- 評価尺度の研究が2本
  - ROUGE-WE(分散表現類似度)の提案@Bloomberg
  - 再評価したらROUGEよりBLUEの方が良い？
- 文生成的(abstractive)要約の研究が3本
  - Attention NNモデルでタイトル生成@Facebook
  - 中国語の短文要約DBを作成、RNN翻訳機を適用
  - 基本意味単位(BSU)で意味ネットワークを構築

全体感(3/3)

- 文抽出的(extractive)要約の研究が13本
  - 劣モジュラ(or Greedy)近似4本
    - 分散表現の分布に基づく劣モジュラ最適化
    - 意味空間の凸包容量のGreedy最大化
    - 主観表現を考慮した劣モジュラ最適化で意見要約
    - フレーズ翻訳と劣モジュラ最適化による言語横断要約
  - その他(ILP、Graph、Tree、EM、LCSなど)
    - URLを含むツイートの生成を抽出的要約として捉える
    - 引用文献の内容を使った論文要約
    - コンセプトベースのILP要約をGreedy近似
    - 頻出アイテムセットを使ったDB圧縮(Krimp)
    - LDA+LexRank
    - RTツリーからCRFで良いツイートを抽出
    - 複数要約アルゴリズムの組み合わせ方
    - EMクラスタリングでPubMed文献のテーマ抽出
    - E-learningでの生徒の反応を要約

- 要約研究の全体感
- 劣モジュラ最適化
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

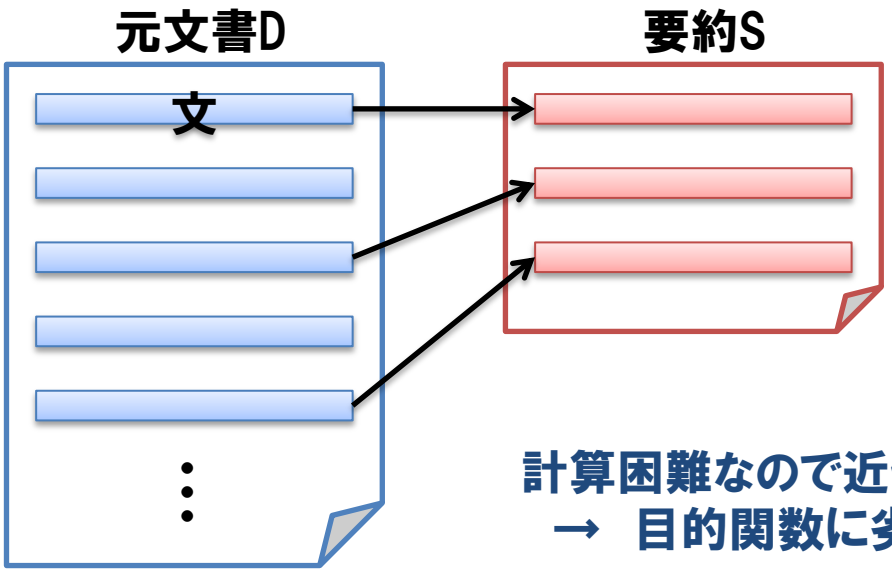
文書要約の定式化

- 文書要約を重要文抽出問題として定式化
  - 元文書の文集合Dから部分集合Sを抽出する問題

$$\max_{S \subset D} f(S) \quad \text{s.t.} \quad c(S) \leq \ell$$

要約の良さを表す関数

文字数制限など



直接的な応用
- ツイートのまとめ作成
- 知恵袋回答のスニペット

間接的な応用
- ニュース要約の前処理
- 商品推薦のスコア

計算困難なので近似したい
→ 目的関数に劣モジュラ性があれば簡単

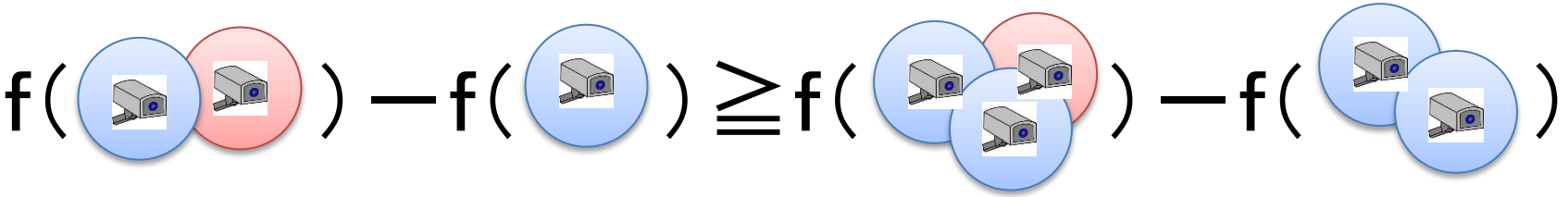
劣モジュラ性

- 連続関数の凸性に対応する集合関数の性質
- 貪欲法でほぼ最適 (1−1/e) 近似が得られる
- [定義] 集合関数$f: 2^X \to \mathbb{R}$が劣モジュラ ⇔

集合S_1, S_2 $(S_1 \subset S_2 \subset X)$と要素$x \in X \setminus S_2$について、

$$f(S_1 \cup \{x\}) - f(S_1) \geq f(S_2 \cup \{x\}) - f(S_2)$$

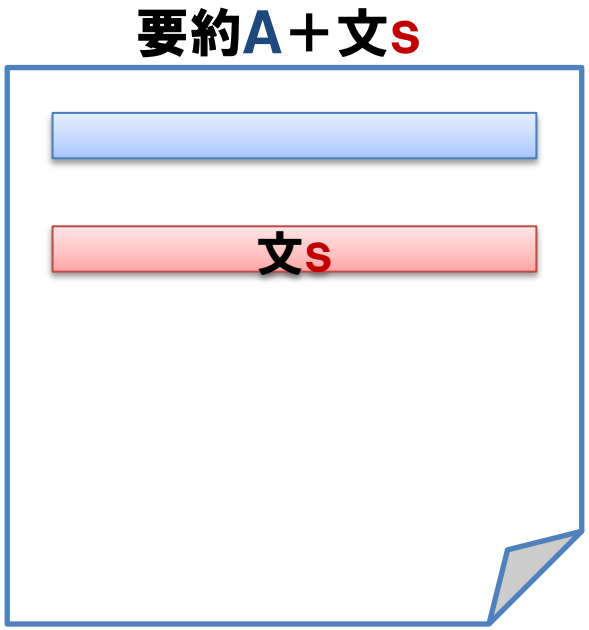
例： センサー配置問題（監視範囲の最大化）

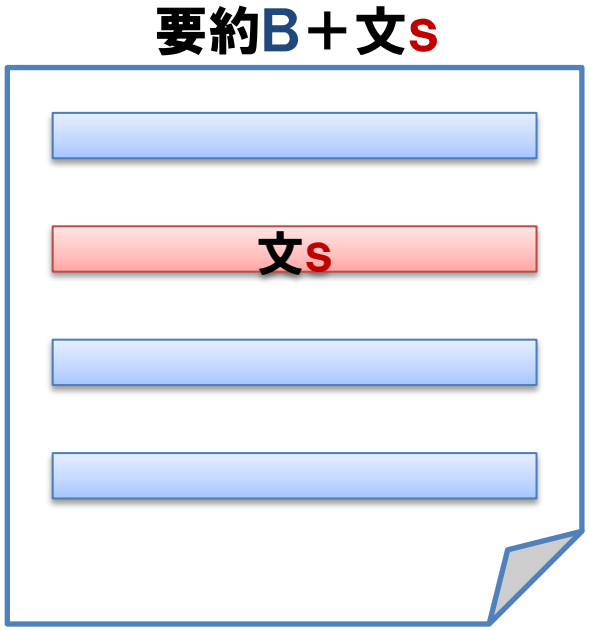


- **元文書の内容を網羅したい→劣モジュラ**

要約A ⊂ 要約B ⇒

$f(A \cup \{s\}) - f(A) \geqq f(B \cup \{s\}) - f(B)$





修正貪欲法 [Lin&Bilmes, ACL2010]

- **要素のコストを考慮した貪欲法**
  - **スコアの増分が大きく、コストが小さい要素を選択**

Algorithm 1: Modified greedy algorithm.

Data: Document D, objective function f, and summary size ℓ.

Result: Summary $C \subset D$.

1 $C \leftarrow \emptyset$; $U \leftarrow D$;

2 **while** $U \neq \emptyset$ **do**

3 $s^* \leftarrow \mathrm{argmax}_{s \in U} f_C(s)/(w_s)^r$;

4 **if** $\sum_{s \in C} w_s + w_{s^*} \leq \ell$ **then** $C \leftarrow C \cup \{s^*\}$;

5 $U \leftarrow U \setminus \{s^*\}$;

6 $s^* \leftarrow \mathrm{argmax}_{s \in D: w_s \leq \ell} f(\{s\})$;

7 **return** $C \leftarrow \mathrm{argmax}_{C' \in \{C, \{s^*\}\}} f(C')$;

$f_C(s) \coloneqq f(C \cup \{s\}) - f(C)$

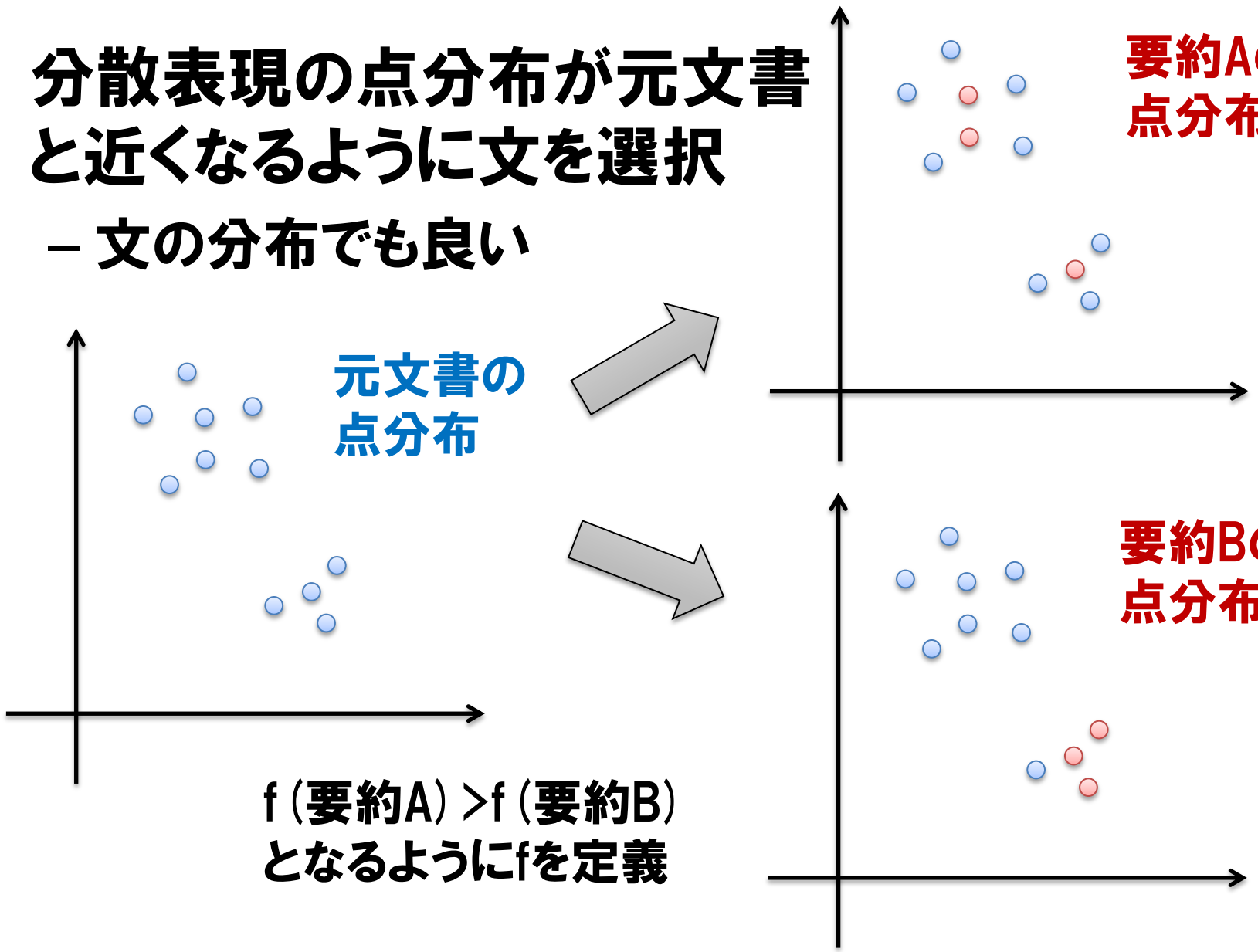
w_sは要素sの重み
（単語数、バイト数など）

近似精度は½(1-1/e)
[Morita+, ACL2013]

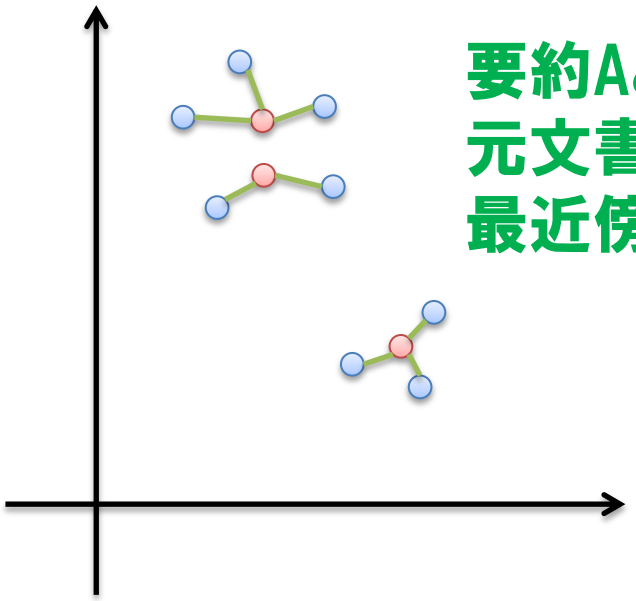
- 要約研究の全体感
- 劣モジュラ最適化
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

- 論文
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
- 内容
  - 分散表現の分布に基づく劣モジュラ関数を提案
  - KL情報量やEarth Mover’s Distanceと関係
- 補足
  - 文書ベクトルの類似度は劣モジュラではない

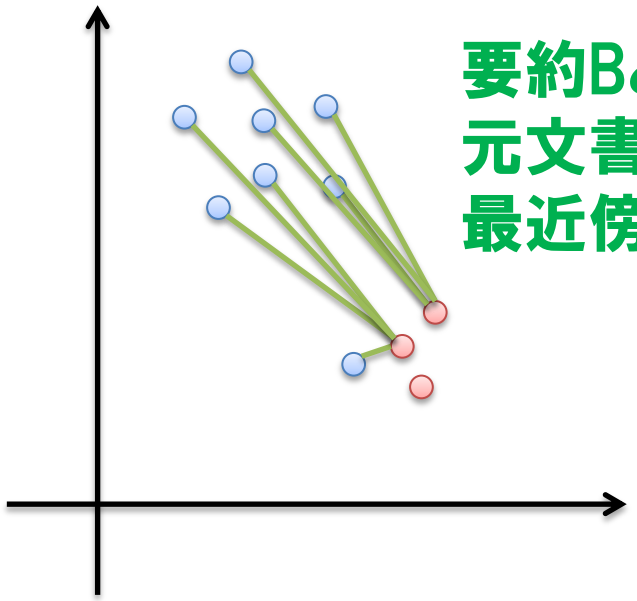
- 分散表現の点分布が元文書と近くなるように文を選択
  - 文の分布でも良い



- 直感:分布が似ている⇒近傍点が近くにある



要約Aと
元文書の
最近傍距離



要約Bと
元文書の
最近傍距離

最近傍点までの距離の(負の)和でfを定義する

- **元文書分布の各点における、要約分布上の最近傍点までの距離の和で非類似度を表す**

$$f^{NN}(C) \coloneqq -\sum_{s \in D} \sum_{w \in s} g(N(w, C))$$

関数gは単調非減少な距離のスケーリング関数

$$N(w, C) \coloneqq \min_{\substack{v \in s : s \in C \\ \vec{w} \neq \vec{v}}} d(\vec{w}, \vec{v})$$

関数Nは単語wからの要約C中の最近傍距離

定理2. f^{NN}**は単調劣モジュラ関数である**

定理3. $g(x) = \ln x$**のとき**f^{NN}**の大小は漸近的にKLDと一致する**

元文書D**、要約**C_1, C_2**について、** $D \sim p, C_1 \sim q, C_2 \sim r$ **とすると漸近的に**

$$\mathbb{E}[f^{NN}(C_2)] - \mathbb{E}[f^{NN}(C_1)] > 0 \Leftrightarrow D_{KL}(p \parallel q) - D_{KL}(p \parallel q) > 0$$

([Perez-Cruz, NIPS2009] [Wang+, TIT2009] などを使う)

データセットと評価指標

- Opinosis Dataset [Ganesan+, COLING2010]
  - 51トピック(ホテル、車、製品など)のユーザレビュー
  - 各トピックに50～575文
  - 各トピックに4，5人が作ったサマリ(１～３文）
- ROUGE-N指標 [Lin, WAS2004]
  - 人が作ったサマリとのNグラム共起割合
  - 翻訳の評価で使われるBLEUに似た評価値
    - BLEUは適合率重視、ROUGEは再現率重視
  - ROUGE-1が最も人のサマリと当てはまりが良い
    - [Lin&Hovy, NAACL2003]

実験結果

- DocEmb: 修正貪欲法＋ f^{Cos}（文書ベクトル）
- EmbDist: 修正貪欲法＋ f^{NN}（点分布）s.t. $g(x) = \ln(x), x, e^x$
- SemEmb: [Kageback et al. CVSC2014]
- TfIdf: [Lin and Bilmes, ACL2011]
- ApxOpt: 修正貪欲法＋ROUGE-1

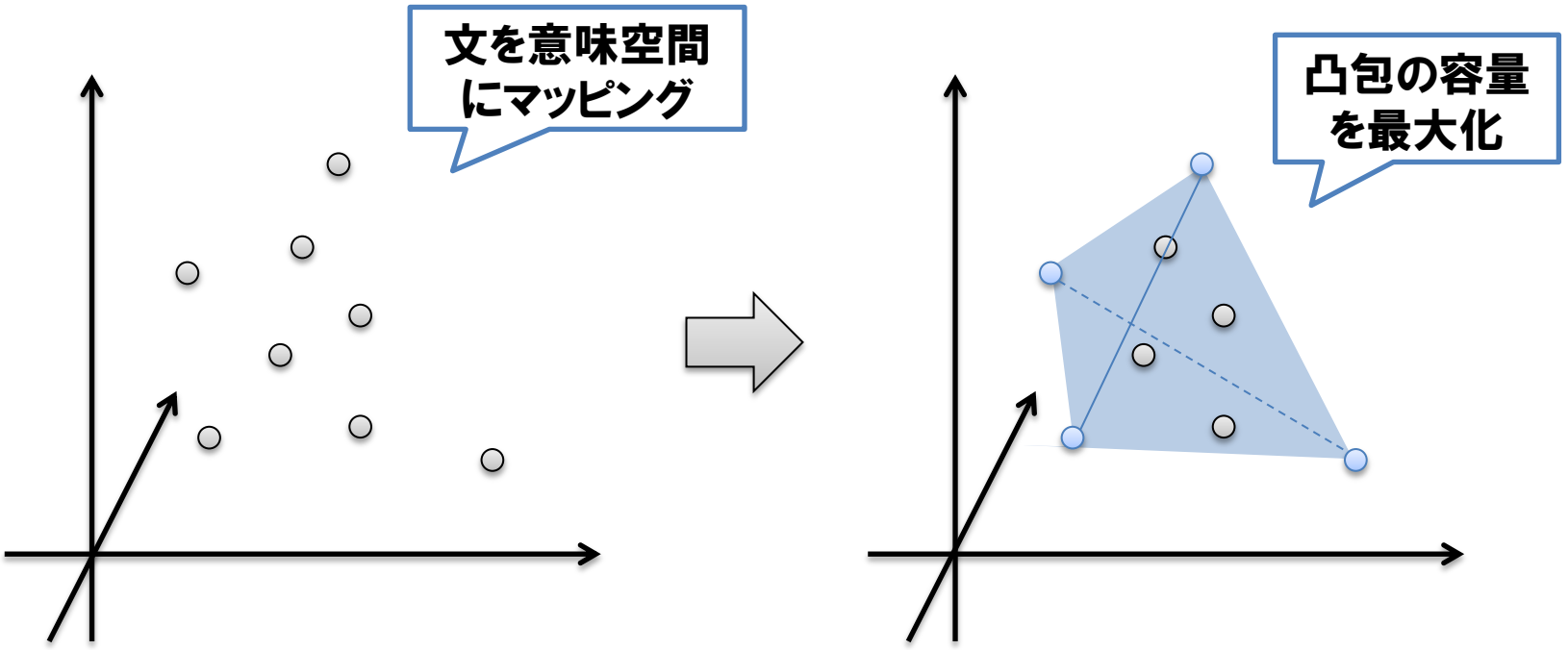
	R-1	R-2	R-3	R-4	
ApxOpt	62.22	21.60	8.71	4.56	近似最適解
EmbDist ($\ln x$)	56.00	16.70	4.93	**1.89**	提案法
EmbDist (x)	55.70	15.73	4.59	1.84	
EmbDist (e^x)	**56.29**	15.96	4.43	1.39	
DocEmb	55.80	13.59	3.23	0.90	コサイン類似度
SenEmb	53.96	15.42	3.97	1.10	既存手法
TfIdf	52.97	**17.24**	**5.40**	1.49	

提案法が最も適した評価指標ROUGE-1で最高性能

- 要約研究の全体感
- 劣モジュラ最適化
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

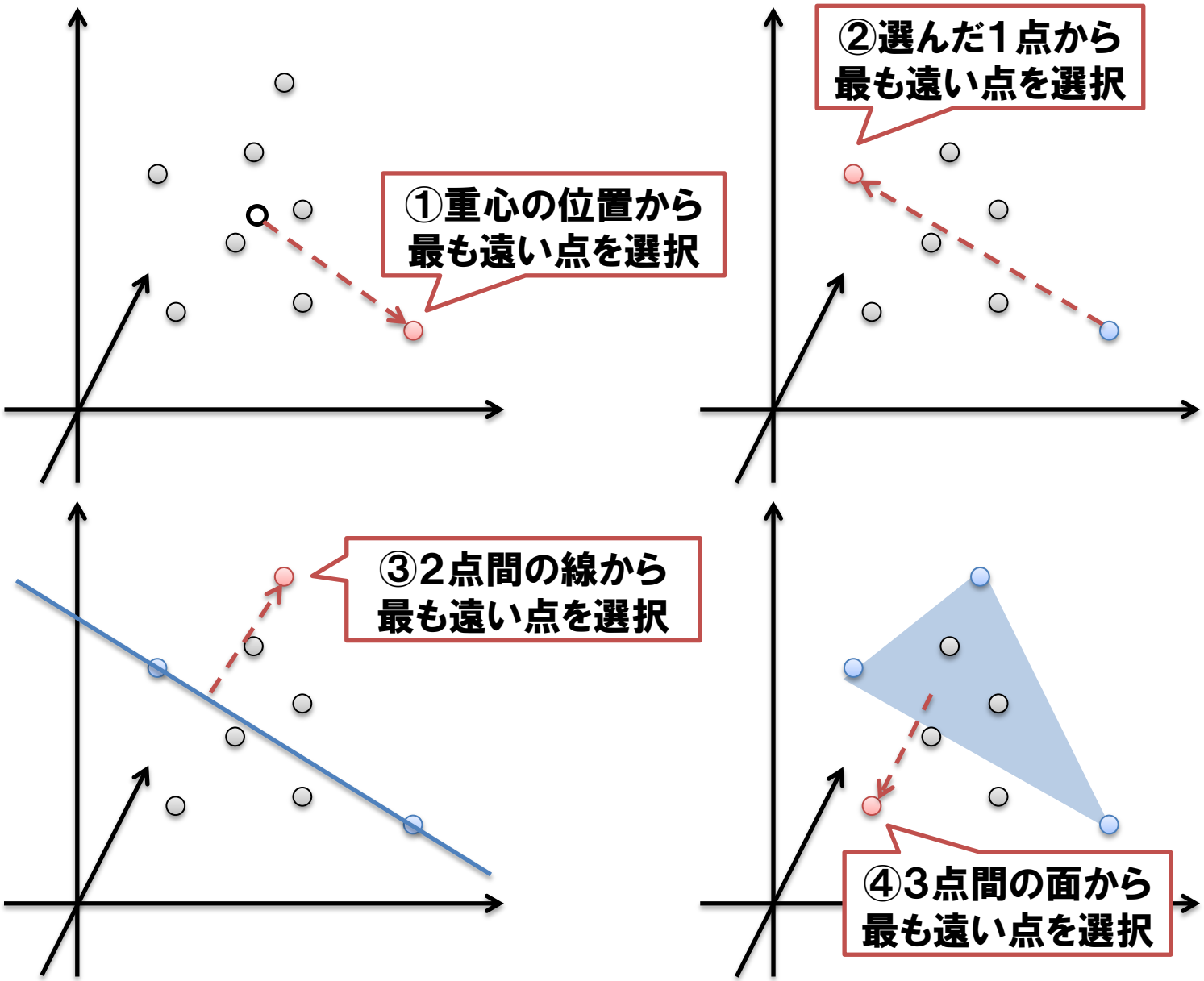
- 論文
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
- 内容
  - 各文を意味空間に埋め込み、文集合によって形作られる凸包の容量を最大化するように文を選択
- 補足
  - Yogatamaさんは現在BaiduのNgグループ
  - 論文のアルゴリズムが間違っている？ようなのでポスターで聞いたアルゴリズムを紹介

提案手法の概要



計算困難なのでGreedyなアルゴリズムで近似
論文には劣モジュラ性についての言及はないが、
上手く目的関数を構成すれば劣モジュラ性を証明できそう

Greedyアルゴリズム



部分空間との距離の計算

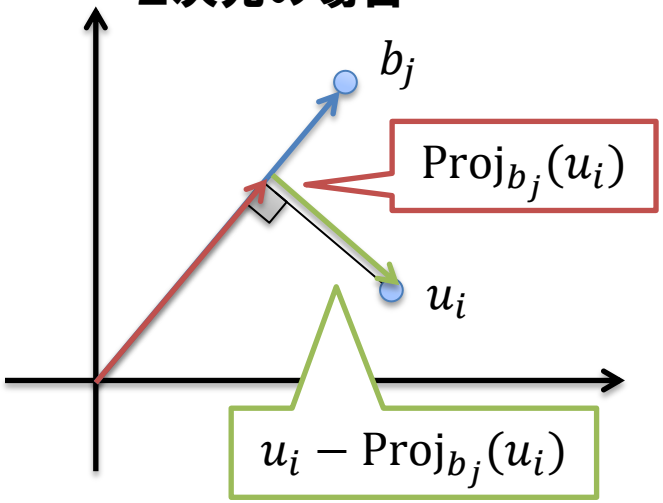
- **部分空間Bとベクトルuの距離**

$$\mathrm{Distance}(\mathbf{u}_i, \mathcal{B}) = \left\| \mathbf{u}_i - \sum_{\mathbf{b}_j \in \mathcal{B}} \mathrm{Proj}_{\mathbf{b}_j}(\mathbf{u}_i) \right\|$$

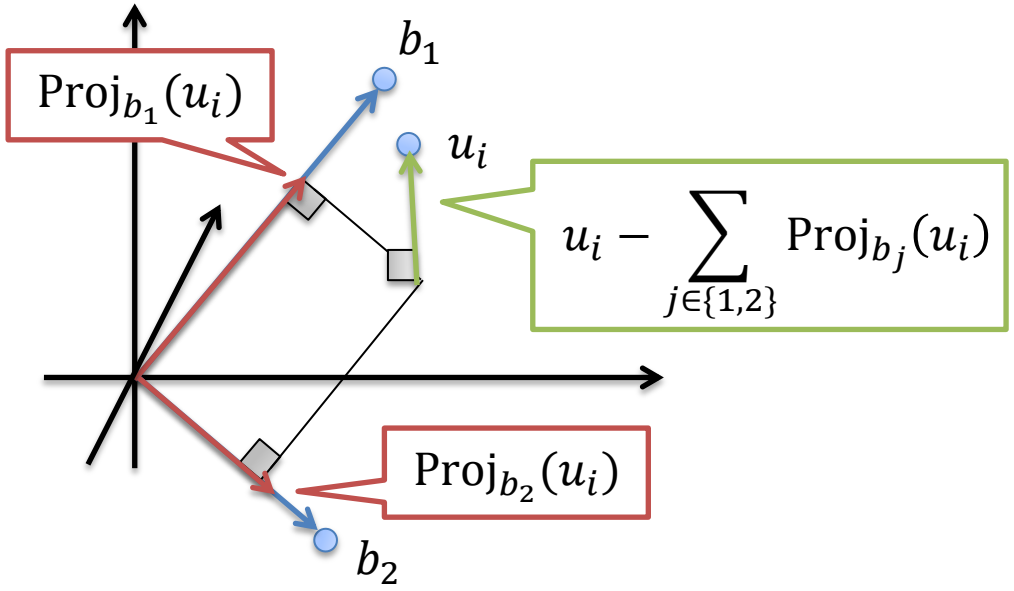
基底ベクトルの集合

Gram-Schmidtの
正規直交化法の式

2次元の場合



3次元の場合



- MMR: 元文書と近いが選択文とは遠い文を選択
  - [Carbonell&Goldstein , SIGIR1998]
- CBS: 2グラムのカバー率を最大化
  - [Gillick+, TAC2008]
- VolumeX: 提案手法＋X次元の空間(SVD)

Methods	TAC-2008		TAC-2009	
	R-1	R-2	R-1	R-2
MMR	34.08	9.30	31.87	7.99
CBS	35.83	9.43	32.70	8.84
Volume 500	37.40	9.17	34.08	**8.91**
Volume 600	**37.50**	**9.58**	**34.37**	8.76
Oracle	46.06	19.33	46.77	16.99

- 要約研究の全体感
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

論文3

- 論文
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
- 内容
  - 意見要約のための劣モジュラ関数を複数提案
  - 内容表現と主観表現のトレードオフを考慮
- 補足
  - 部分列挙と組み合わせたGreedyを利用

提案手法の概要

- 内容表現と主観表現のトレードオフを表した劣モジュラ関数を最適化して要約
  - ベースは [Lin&Bilmes, ACL2011]

$$F(S) = \alpha L(S) + \beta A(S)$$

内容表現スコア
要約候補Sと元文書Vの類似度で定義

$$L(S) = \sum_{i \in V} min\{c_i(S), \gamma c_i(V)\}$$

$$c_i(S) = \sum_{j \in S} w_{i,j}$$

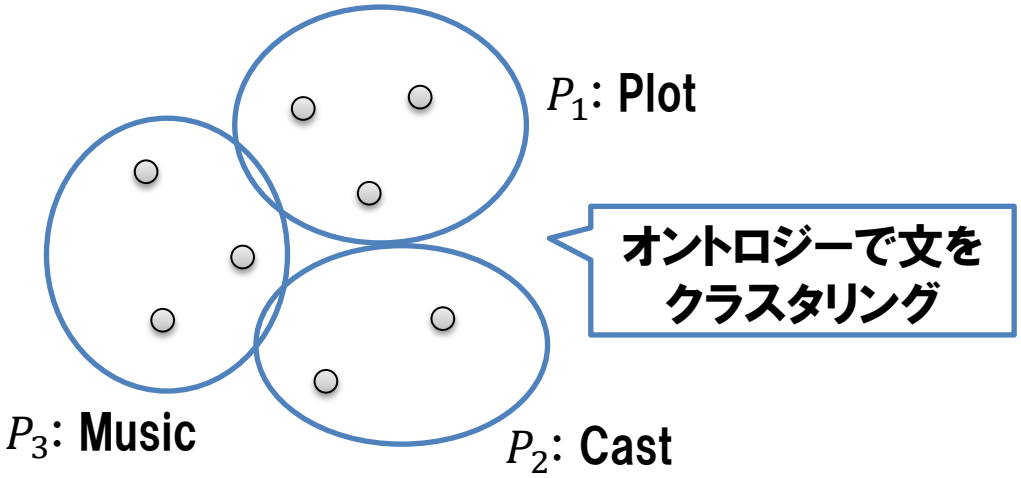
$w_{i,j}$**は文iと文jの類似度**

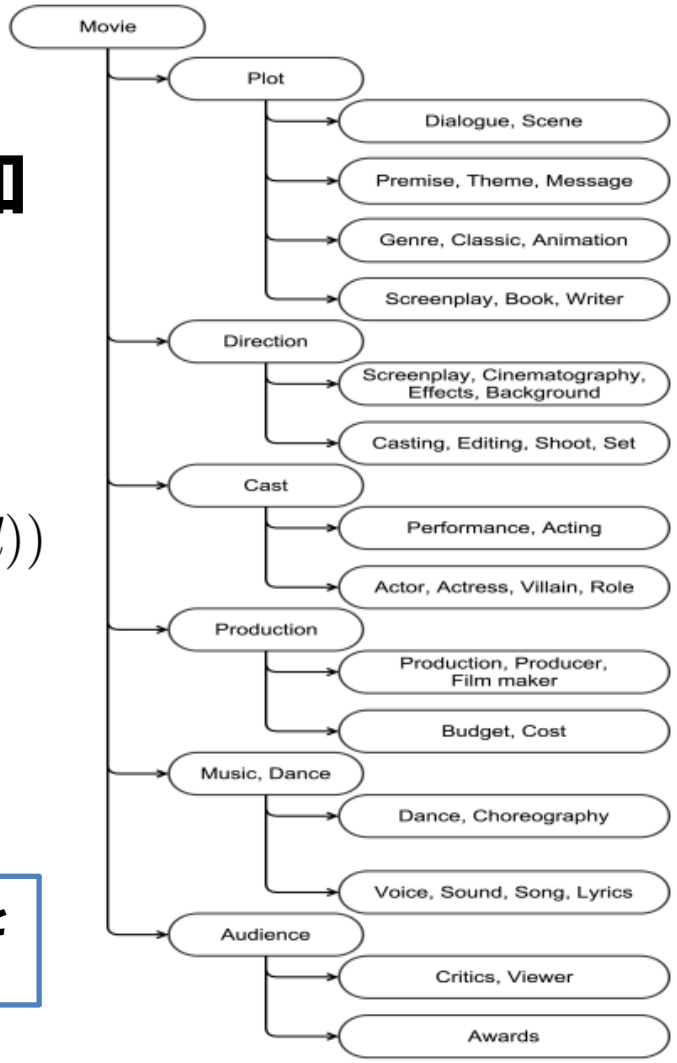
主観表現スコア
主観表現を多く含む文に高いスコア
論文では5種類提案

- A1: Modular Function
  - 各評価軸のPN値の重み付き和

$$A_1(S) = \sum_i \sum_{j \in (P_i \cap S)} s_j * w_i$$

$$s_j = \sum_{word \in j} (pos(word) + neg(word))$$





- A2: Badget-additive Function
  - 評価軸ごとに閾値を設けて多様性を確保

$$A_2(S) = \sum_i min(\sum_{j \in (P_i \cap S)} s_j, \lambda_i) * w_i$$

- A3: Polarity Partitioned Budget-additive Func.
  - P/Nも独立に閾値を設けてP/Nの多様性も確保

$$A_3(S) = \sum_i min(\sum_{j \in (P_i \cap S \cap P_{pos})} s_j, \lambda_i) * w_i + min(\sum_{j \in (P_i \cap S \cap P_{neg})} s_j, \lambda_i) * w_i$$

- A4: Facility Location Function
  - 各評価軸の最大PN値の重み付き和

$$A_4(S) = \sum_i max_{j \in (P_i \cap S)} s_j * w_i$$

- A5: Polarity Partitined Facility Location Func.
  - P/Nを独立に扱い多様性を確保

$$A_5(S) = \sum_i max_{j \in (P_i \cap S \cap P_{pos})} s_j * w_i + \sum_i max_{j \in (P_i \cap S \cap P_{neg})} s_j * w_i$$

実験結果

- TOP: 先頭X文
- TOP-SUBJ: 主観表現スコア上位X文
- LER-SM: 元文書のPNと近い文を選択 [Lerman+, EACL2009]
- TEXTRANK: 文の類似度グラフをPageRank [Mihalcea&Tarau, EMNLP'04]
- MINCUT: 最大流問題として主観/客観分類 [Pang&Lee, ACL2004]

System	ROUGE1	ROUGE2	S. Corr.
TOP	0.43001	0.16591	0.86144
TOP-SUBJ	0.41807	0.14362	0.82953
LER-SM	0.42608	0.14533	0.96545
TEXTRANK	0.41987	0.14644	0.88967
MINCUT	0.39368	0.11047	0.84017
Submod-A_1	0.43223	0.15702	0.95306
Submod-A_2	0.43594	0.15977	0.97538
Submod-A_3	0.43247	0.15436	0.93155
Submod-A_4	**0.43602**	**0.15760**	**0.98566**
Submod-A_5	0.42976	0.15551	0.95415

Naïve Bayesによる元文書のPN判定と要約のPN判定の関連度

提案手法

（映画レビュー [Pang&Lee, ACL2004] ＋人手サマリを用いた）

- 要約研究の全体感
- 劣モジュラ最適化
- 劣モジュラ関係の論文を4本紹介
  - Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka, “Summarization Based on Embedding Distributions”, EMNLP 2015
  - Dani Yogatama, Fei Liu, Noah A. Smith, “Extractive Summarization by Maximizing Semantic Volume”, EMNLP 2015
  - Jayanth Jayanth, Jayaprakash Sundararaj, Pushpak Bhattacharyya, “Monotone Submodularity in Opinion Summaries”, EMNLP 2015
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015

- 論文
  - Jin-ge Yao, Xiaojun Wan, Jianguo Xiao, “Phrase-based Compressive Cross-Language Summarization”, EMNLP 2015
- 内容
  - フレーズ翻訳の応用で言語横断要約
  - 圧縮(翻訳)した文をGreedy選択
- 補足
  - 実際には文抽出的手法ではない
  - フレーズ単位の翻訳とアライメントはできると仮定

提案手法の概要

- フレーズベース翻訳の目的関数を転用して、文圧縮(翻訳)した後に文選択する
  - フレーズベース翻訳の目的関数

($y = (p_1, \ldots, p_L)$はフレーズ列で表された元言語の文書)

$$f(y) = \sum_{k=1}^{L} g(p_k) + LM(e(y)) + \sum_{k=1}^{L-1} \eta |start(p_{k+1}) - 1 - end(p_k)|$$

フレーズの重要度 / 目標言語の自然さ / 並び替えのペナルティ

  - フレーズベース文選択の目的関数

$$F(S) = \sum_{p \in S} \sum_{i=1}^{count(p,S)} d^{i-1} g(p) + \sum_{s \in S} bg(s) + \eta \sum_{s \in S} dist(y(s))$$

フレーズの重要度 / 2gramスコアLMの代用 / 並び替えのペナルティ

減衰率dをかけて同じフレーズを選ばない

翻訳が難しい文のスコアを下げる

- **類似のスコア関数でフレーズ抽出して結合**

Algorithm 2 A growing algorithm for finding the maximum density compressed sentence

```
function GET_MAX_DENSITY_COMPRESSION(s, S_{i-1})
    queue Q ← ∅, kept ← ∅
    for each phrase p in s.phrases do
        if p.score/p.cost > 1 then
            kept ← kept ∪{p}
            Q.enqueue(p)
        end if
    end for
    while Q ≠ ∅ do
        p ← Q.deque()
        ppv ← p.previous_phrase, pnx ← p.next_phrase
        if (ppv.score+bg(ppv,p)+ηdist(ppv,p)) / (ppv.cost+p.cost) > 1 then
            Q.enqueue(ppv), kept ← kept ∪{ppv}
        end if
        if (pnx.score+bg(pnx,p)+ηdist(p,pnx)) / (p.cost+pnx.cost) > 1 then
            Q.enqueue(pnx), kept ← kept ∪{pnx}
        end if
    end while
return s̃ = kept, ratio = (F(S_{i-1}∪{s̃}) − F(S_{i-1})) / s̃.cost
end function
```

①フレーズ密度が高いものを選択

密度＝ $\dfrac{\sum_{i=1}^{count(p,S_{i-1})} d^{i-1} g(p)}{|p|}$

②選択されたフレーズの前後から結合して密度が高いものを選択

結合したフレーズ列をqとすると、

密度＝ $\dfrac{\text{qの文選択スコア}}{|q|}$

- Baseline (EN) : 文抽出後に翻訳
- Baseline (CN) : 翻訳後に文抽出
- PBES: 提案法(圧縮なし)
- CoRank: [Wan, ACL2011]
- Baseline (ENcomp) : 文抽出・圧縮後に翻訳
- PBCS: 提案法

Character Budgeting	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-L	ROUGE-SU4
Baseline(EN)	0.21460	0.03494	0.05150	0.12343	0.06278
Baseline(CN)	0.21589	0.03732	0.05420	0.12867	0.06405
PBES	0.22825	0.04037	0.05527	0.12856	0.06894
CoRank (reimplemented)	0.22593	0.04069	0.05887	0.12818	0.07241
Baseline(ENcomp)	0.23663	0.04245	0.06134	0.13070	0.07365
PBCS	**0.24917**	**0.04632**	**0.06252**	**0.13591**	**0.07953**

Table 1: Results of word-based ROUGE evaluation

- ご清聴ありがとうございました！

EMNLP2015会場の様子（リスボン）