# アスペクトの被覆を実現するための最小値最大化問題に基づく文書要約モデル

牧野 拓哉[1]　　高村 大也[2]　　奥村 学[2]
1 東京工業大学 総合理工学研究科
2 東京工業大学 精密工学研究所
makino@lr.pi.titech.ac.jp, {takamura, oku}@pi.titech.ac.jp

## 1　はじめに

本論文では，文書要約タスクの一つである guided summarization タスクにおいて，最大被覆問題と最小値最大化問題を組み合わせた新たな文書要約モデルを提案する．guided summarization タスクでは，含めるべき情報にアスペクトという指向性がある．アスペクトは読者が元の文書集合の内容を知るために重要な項目である．generic summarization タスクで良い結果を出している最大被覆問題に基づく文書要約モデルは，文を概念単位の集合として表現し，できるだけ多くの重要な概念単位を被覆するように要約を生成する．概念単位の重みは，文書頻度を利用して計算されるが，guided summarization では，アスペクトを被覆する必要があるため，文書頻度を利用した重みの計算方法では目的を達成することができない．そこで，本論文では，アスペクトのバランスの取れた被覆を最小値最大化問題として定式化する．最小値最大化問題は，要約に含まれるアスペクトの反映度のうち，最小値を最大化するため，バランスの良いアスペクトの被覆ができる．文書要約の評価手法としてよく用いられる ROUGE を用いて，他の手法との比較，評価をおこなう．

## 2　関連研究

近年，文書要約の分野で整数線形計画問題 (Integer Linear Programming Problem (ILP)) を利用した手法が盛んに研究されている．ILP は，変数の取りうる値が整数であるような制約がついた線形計画問題で表される．文書要約問題は長さの制約の範囲内で元の文書クラスタの情報を最大化するような問題として表現されることが多い．ここで要約技術の背景について簡単に述べておく．まず Carbonell and Goldstein[1] は，文の逐次選択による文書要約を実現した．彼らは，すでに選択された文と似た文に対してペナルティを与えることで要約における冗長性を排除した．ここで使われるペナルティは maximal marginal relevance (MMR) と呼ばれる．McDonald[2] は，ILP として文書要約を表現し，動的計画法を用いて解いた．

Takamura and Okumura[3] は，多様性に加えて与えられた文書クラスタの主題への関連性を考慮した文書要約モデルへの拡張をおこなっている．このモデルでは，文を概念単位の集合と考える．概念単位について，文書頻度に比例した重要度を与えたり，logistic regression を用いて概念単位が人手の要約に入る確率を計算 [4] し，それを重要度として与え，より多くの重要な概念単位を被覆するように文を選択して要約を生成することが目的となる．さらに，彼らは，要約にある程度冗長性を持たせることによって，要約の結束性や一貫性を表現した．できるだけ多くの概念単位を被覆することと，要約に結束性，一貫性を持たせることを同時に表現したより直感的な文書要約モデルを提案した．

## 3　提案手法

### 3.1　各文のアスペクトの反映度の計算方法

我々は，文書クラスタ中のアスペクトを捉えるために最大エントロピー分類器を用いた．最大エントロピー分類器の出力は，文 $j$ がどの程度の確率でクラス $a$ に所属しているかを表しており，我々はこの値を文がどの程度アスペクト $a$ を反映しているかを表す値とみなす．提案手法では抽出型の要約をおこなうため，文単位でアスペクトの反映度の予測をおこなう．今回

用いた素性値を式 (1) に示す:

$$\phi_k(j,y) = \begin{cases} 1, & \text{if n-gram } k \text{ appears in } j \text{ and } y = a \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

素性値 $\phi_k(j,y)$ は，文 $j$ が n-gram $k$ を含み，かつラベル $y$ が $a$ であった ときに 1 となり，それ以外は 0 となる．n-gram には訓練データ中の unigram と bigram を用いた．文 $j$ におけるアスペクト $a$ の反映度は式 (2) のように表される:

$$p(y|j) = \frac{1}{Z(j)} \exp\left( \sum_k {}_k\phi_k(j,y) \right). \quad (2)$$

ただし，$Z(j)$ は正規化定数で，${}_k$ は素性 $k$ の重みである．訓練データには TAC から配布されている正解データ (pyramid ファイル) を用いた．pyramid ファイルではアスペクトのラベルが文以下の粒度のテキストに対して付けられているが，我々は文を単位としてアスペクトの反映度を予測するので，ラベルが付けられたテキストを含む文が，そのアスペクトを反映しているものと考える．最大エントロピー分類器はアスペクトを反映しているか，そうでないかの二値分類をおこなうため，各アスペクトについて最大エントロピー分類器を訓練する．正例は，アスペクト $a$ とラベルが付けられた文，負例は $a$ 以外のアスペクトのラベルが付けられた文として二値分類用の訓練データを用意する．一つの文で複数のアスペクトを反映している場合があるが，訓練データ中に正例かつ負例にも現れる文があるような場合には，負例の文を削除している．

## 3.2 アスペクト被覆のためのモデル化

提案手法では，彼らのモデルに対してアスペクトの被覆を可能にするための項を線形結合したモデル化をおこなう:

$$\begin{aligned} \text{max.} \quad & (1-\beta)\left\{ \alpha \sum_i w_i c_i + (1-\alpha) \right. \\ & \left. \sum_j \left( \sum_i w_i o_{ij} \right) s_j \right\} + \beta z, \quad (3) \\ \text{s. t.} \quad & \\ & \forall j, \sum_j s_j o_{ij} \quad c_i; \quad \forall i, \forall j, s_j o_{ij} \le c_i, \\ & \sum_j s_j l_j \le L, \\ & \forall a \in aspects, \sum_j s_j p_{aj} \quad z, \\ & \forall i, c_i \in \{0,1\}; \quad \forall j, s_j \in \{0,1\}. \end{aligned}$$

ここで $c_i$ を，要約が概念単位 $i$ を含むときに 1，それ以外の場合は 0 となるような変数とする．また $s_j$ を，要約が文 $j$ を含むときに 1, それ以外の場合は 0 となるような変数にする，さらに，$o_{ij}$ を文 $j$ が概念単位 $i$ を含むときに 1，それ以外の場合は 0 となるような定数とする．$w_i$ は概念単位 $i$ が持つ重要度とする．式 (3) の初項，第二項はそれぞれ多様性，文書クラスタが持つ主題との関連性を表している．この二つの項をパラメータ $\alpha$ で線形結合する．$\alpha$ は，$[0,1]$ の実数値で，値が大きいほど多様性を重視し，値が小さいほど主題との関連性を重視するパラメータである． 本論文で提案しているアスペクトの被覆をおこなうための項が第三項である．$z$ という変数は，要約のアスペクトの中で最も低い反映度を表している．要約が持つ任意のアスペクト $a$ の反映度は，要約に含まれる文集合のアスペクト $a$ の反映度の総和 $\sum_j s_j p_{aj}$ としている．つまりこの変数 $z$ を最大化することで，反映度が最も低いアスペクトのスコアを最大化することになるので，網羅的にアスペクトの被覆をおこなうことをモデル化していることになる．要約の長さ制約 $L$ の範囲内で文を選択して要約を生成する．ただし文 $j$ の長さは $l_j$ で，ここでは長さ制約 $L$ も各文の長さ $l_j$ も単語数とする．

最大エントロピー分類器による反映度の予測ではすべての文に実数値が割り当てられるが，文書要約問題を解くときに用いる反映度は各アスペクトについて上位 $N_a$% までとする．これは反映度が高いわけではないが，複数のアスペクトである程度の反映度を割り当てられ，結果として不当に要約のアスペクトの反映度を高くしてしまうような文を要約へ含めることを避けるためである．Takamura and Okumura の提案した多様性と主題との関連性の 2 つの項と，アスペクトの被覆を可能にするための項をパラメータ $\beta$ で線形結合する．$\beta$ は，$[0,1]$ の実数値で，値が大きいほどアスペクトの被覆を重視し，値が小さいほどアスペクトの被覆は重視しない．つまり，$\beta = 0$ のときに，Takamura and Okumura のモデルと一致する．

提案手法として比較する手法として，以下の定式化をおこなう．この定式化は，網羅的なアスペクトの被覆をおこなおうとはせず，アスペクトの種類を問わずにスコアが高い文が選択されやすくなる:

$$\begin{aligned} \text{max.} \quad & (1-\beta)\left\{ \alpha \sum_i w_i c_i + (1-\alpha) \right. \\ & \left. \sum_j \left( \sum_i w_i o_{ij} \right) s_j \right\} + \beta \sum_a z_a, \quad (4) \\ \text{s. t.} \quad & \end{aligned}$$

$$\forall j, \sum_j s_j o_{ij} \quad c_i; \quad \forall i, \forall j, s_j o_{ij} \leq c_i,$$
$$\sum_j s_j l_j \leq L,$$
$$\forall a \in aspects, \sum_j s_j p_{aj} = z_a,$$
$$\forall i, c_i \in \{0,1\}; \quad \forall j, s_j \in \{0,1\}.$$

# 4 実験及び考察

## 4.1 実験設定

テストデータには TAC 2010 のデータセットと TAC 2011 のデータセットを用いた．TAC 2010 と TAC 2011 における複数文書要約タスクにおいて，参加者にはそれぞれ 44，46 個の文書クラスタが与えられる．文書クラスタは同じ事象について述べている 10 の新聞記事から構成されていて，100 単語以内で要約を生成することが求められる．また，すべての文書クラスタは 5 つのトピックのいずれか 1 つに必ず属しており，各トピックには平均して 6 つのアスペクトが予め定義されている．ただし，文書クラスタがどのトピックに属しているかは参加者に予め知らされている．圧縮率は TAC 2010，TAC 2011 ともに平均で 2% である．評価には ROUGE[5] を用いた．本論文では，ROUGE-2 を用いる．有意差検定には Wilcoxon の符号順位検定を用いた．有意水準は 0.05 に固定した．ただし，トピックごとの文書クラスタは数が少ないため，すべてのトピックをまとめて有意差検定をおこなう．最大エントロピー分類器の訓練データは，2010 年の文書クラスタに対しては 2011 年の pyramid ファイルを，2011 年の文書クラスタに対しては 2010 年の pyramid ファイルをそれぞれ用いる．パラメータ $\alpha$，$\beta$ については，それぞれ 0 から 1 まで 0.1 刻みで変化させ，TAC 2010, TAC 2011 における同じトピックに属するテストデータ以外の文書クラスタを訓練データとして，訓練データの ROUGE-2 値が最大となるように決定した．

予備実験により，本論文では概念単位は bigram とし，Takamura and Okumura のモデルの ROUGE-2 値が高くなるように，重みが大きい上位 600 個の概念単位を用いる．また，概念単位の重みは文書頻度に比例した重みを付け，$N_a = 30\%$ とした．Yih et al. や Takamura and Okumura では，logistic regression を訓練して概念単位の重みを計算した場合のほうが文書頻度に比例した重みをつけるよりも ROUGE 値が高いと報告されている．本論文でも，TAC の参照要約を用いて logistic regression を訓練すると間接的にアスペクトを考慮した重みが計算できると考えたが，訓練データが十分に多くないため，我々の実験では logistic regression を訓練して計算した重みは用いない．

## 4.2 実験結果

表 1，表 2 に結果を示す．*baseline* の ROUGE-2 値に対して有意に高い ROUGE-2 値には下線を引いてある．*baseline* は Takamura and Okumura の最大被覆問題に基づく文書要約モデルである．*balanced* は式 (3) で表される提案手法で，*not-balanced* は式 (4) で表される文書要約モデルとなっている．また，*peer 22*[6], *peer 43* はそれぞれの年で最も ROUGE-2 値が高かったシステムである．ただし，*peer 22*, *peer 43* は，他の手法と比べて訓練データが無い，あるいは半分であるため，これらの結果と提案手法との公平な比較にはなっていない．括弧の中の値域は 95[%] 信頼区間を表す．表 1 では，*balanced* と *not-balanced* の ROUGE-2 値がともに *baseline* の ROUGE-2 値を有意に上回っているが，表 2 では，*not-balanced* は *baseline* の ROUGE-2 値を有意に下回っているのに対し，*balanced* は *baseline* の ROUGE-2 値を有意に上回っている．表 1 の Topic が Trials & Investigations のとき，表 2 の Topic が Attacks のときに，*balanced* の ROUGE-2 値が，*baseline* の ROUGE-2 値を下回っている．これは，パラメータを決める際の訓練データが十分に多くないことと，最大エントロピー分類器の反映度の予測が正確でないことの二つに起因していると考えられる．*balanced* の ROUGE-2 値が *baseline* の ROUGE-2 値よりも下がっている文書クラスタについて決定されたパラメータ $\beta$ は，同じトピックに属する他の文書クラスタについて決定されたパラメータ $\beta$ よりも高くなっていた．これにより，アスペクトの被覆が概念単位の被覆や，主題との関連性よりも重視されやすくなる．このようなパラメータで，最大エントロピー分類器の反映度の予測が正確でないため，本来アスペクトを反映していないような文が要約として多く選択されて ROUGE-2 値が下がってしまったと考えられる．

次に，提案手法の ROUGE-2 値が *baseline* の ROUGE-2 値よりも向上した例を挙げる．トピック Disasters に属する *1046H-A* について，*baseline* には選択されないが，*balanced* には選択された文に *An extremely powerful earthquake and the tsunami that followed devastated many parts of North Sumatra and Aceh on Dec. 26 last year, killing more than 100,000 people there.* がある．この文は，Disasters で

表 1: TAC 2010 における ROUGE-2．有意差検定は，平均に対してのみおこなっている．

| Topic \ system | *baseline* | *not-balanced* | *balanced* | *peer 22* |
|---|---|---|---|---|
| Disasters | 0.082 (0.061 - 0.102) | 0.089 (0.072 - 0.108) | 0.094 (0.075 - 0.119) | 0.089 (0.066 - 0.112) |
| Attacks | 0.121 (0.089 - 0.150) | 0.125 (0.092 - 0.155) | 0.121 (0.089 - 0.150) | 0.102 (0.076 - 0.131) |
| Health & Safety | 0.071 (0.049 - 0.093) | 0.078 (0.057 - 0.098) | 0.080 (0.056 - 0.102) | 0.062 (0.043 - 0.078) |
| Endangered Resources | 0.096 (0.071 - 0.123) | 0.101 (0.079 - 0.135) | 0.105 (0.082 - 0.129) | 0.101 (0.082 - 0.117) |
| Trials & Investigations | 0.160 (0.137 - 0.180) | 0.154 (0.129 - 0.180) | 0.158 (0.134 - 0.180) | 0.134 (0.109 - 0.158) |
| 平均 | 0.104 (0.091 - 0.119) | 0.109 (0.095 - 0.123) | 0.110 (0.096 - 0.124) | 0.096 (0.084 - 0.107) |

表 2: TAC 2011 における ROUGE-2．有意差検定は，平均に対してのみおこなっている．

| Topic \ system | *baseline* | *not-balanced* | *balanced* | *peer 43* |
|---|---|---|---|---|
| Disasters | 0.139 (0.114 - 0.167) | 0.131 (0.105 - 0.158) | 0.148 (0.122 - 0.174) | 0.149 (0.128 - 0.170) |
| Attacks | 0.162 (0.137 - 0.195) | 0.160 (0.132 - 0.196) | 0.161 (0.136 - 0.194) | 0.153 (0.126 - 0.181) |
| Health & Safety | 0.113 (0.082 - 0.141) | 0.120 (0.088 - 0.150) | 0.120 (0.088 - 0.151) | 0.133 (0.100 - 0.164) |
| Endangered Resources | 0.077 (0.057 - 0.099) | 0.080 (0.056 - 0.106) | 0.079 (0.060 - 0.100) | 0.085 (0.069 - 0.104) |
| Trials & Investigations | 0.132 (0.098 - 0.173) | 0.116 (0.086 - 0.148) | 0.132 (0.098 - 0.173) | 0.148 (0.111 - 0.186) |
| 平均 | 0.126 (0.110 - 0.142) | 0.123 (0.108 - 0.139) | 0.129 (0.113 - 0.145) | 0.134 (0.119 - 0.149) |

定義されているアスペクトのうち，WHAT，WHERE，WHEN，DAMAGES を反映しており，この文だけでも文書クラスタでどのような事象について述べられているのかの概観を理解することができる．このような文に含まれる概念単位の重要度が高くないが，事象の概観をつかみやすい文を提案手法では選択することができた．

## 5 おわりに

バランスの良いアスペクトの被覆を実現させるために最小値最大化問題として定式化し，先行研究のモデルと組み合わせた新たなモデルを提案した．今後の課題としては，アスペクトの反映度を捉えるための最大エントロピー分類器の精度をより向上させたいと考えている．本論文では，unigram，bigram を含んでいるかどうかの二値ベクトルを素性としたシンプルな素性であったので，さらなる工夫が可能である．例えば，non-factoid 型の質問応答タスクにおいて，Basic Element[7]，意味的素性や，文法的素性を用いて文を表現することで質問文と文書中の文の類似度を計算し，回答文を抽出することが有効であることが報告されており [8]，guided summarization タスクにもこのような素性を応用可能である．さらに，アスペクトが考慮された参照要約が十分な量手に入った場合に，logistic regression を用いて概念単位に間接的にアスペクトを考慮した重みを与えた場合にも，最小値最大化問題によってアスペクトを被覆する項が有効であるのか確かめる必要がある．

## 参考文献


[1] Jaime Carbonell and Jade Goldstein: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pp.335–336, (1998).

[2] Ryan McDonald: A study of global inference algorithms in multi-document summarization. In *Proceedings of ECIR*, pp.557–564, (2007).

[3] Hiroya Takamura and Manabu Okumura: Text Summarization Model based on Maximum Coverage Problem and its Variant, In *Proceedings of EACL*, pp.781–789, (2009).

[4] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki: Multi-Document Summarization by Maximizing Informative Content-Words, In *Proceedings of IJCAI*, pp.1776–1782 (2007).

[5] Chin-Yew Lin and Eduard Hovy: Automatic evaluation of summaries using n-gram cooccurrence statistics. In *Proceedings of NAACL*, pp.71–78 (2003).

[6] Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharath Reddy, Sudheer Kovelamudi, Srikanth Reddy Vaddepally, Radheshyam Nanduri, Kiran Kumar N, Santhosh Gsk and Prasad Pingali: IIIT Hyderabad in Guided Summarization and Knowledge Base Population, In *Proceedings of TAC* (2010).

[7] Eduard Hovy, Chin-Yew Lin, Liang Zhou and Junichi Fukumoto: Automated Summarization Evaluation with Basic Elements, In *Proceedings of LREC*, (2006).

[8] Yllias Chali and Sha q Rayhan Joty: Selecting Sentences for Answering Complex Questions, In *Proceedings of EMNLP*, pp.304–313. (2008).