Project Next Summarization:
Project Next 要約タスク最終報告

1 はじめに

Project Next 要約タスクの最終報告では，以下の参加者よりレポートを受け付けた．

- 小倉由佳里 (お茶大): 多目的遺伝的アルゴリズムを用いた複数文書要約
- 鈴木聡子 (お茶大): グラフを用いた時系列文書の要約
- 森田一 (京都大学): Project Next: 文圧縮を利用したクエリ指向要約を対象にした誤り分析
- 高村大也，菊池悠太 (東工大): 単一文書要約における談話構造の効果と課題
- 西川仁 (NTT): Project Next Summarization: 自動要約タスクにおける誤り分析の枠組みの提案
- 平尾努 (NTT): 談話依存構造木に基づく要約手法の誤り分析

2 最終報告作成の方針

中間報告に対するアドバイザの方々のコメントをふまえ，以下に示す西川の 3 つの観点 (「Project Next Summarization: 自動要約タスクにおける誤り分析の枠組みの提案」を参照のこと) に基づき要約システムのエラー分析を進めることを決定した．

1. 文の文法性に関して: 要約システムが非文を出力していないか．
2. 情報の網羅性に関して: 出力から読み取れる情報が，入力および読み手の希望を鑑みて，重要であること．重要でない，枝葉末節の情報が出力に含まれないこと．
3. テキストの一貫性に関して: 読み取れる情報が，入力と矛盾せず，入力が出力を含意すること．読み手が入力を読んだ際と出力を読んだ際に異なる結論に至らないこと．

なお，上記以外の方針は以下のとおり．

- 分析に利用するコーパスは言語も含め分析者に一任する．要約タスクグループとして特定のコーパスを対象とはしない．これは，要約タスクが多様であるため特定のコーパスを分析対象とすることが現実的でないと判断したためである．
- 同様に分析に用いる自動要約システムも分析者に一任する．

3 分析結果の傾向

3.1 文の文法性に関して

森田は文抽出と圧縮を同時に行うクエリ指向の要約システムにおける文圧縮のエラーを以下の 2 段階に分けて分析している．

1. 圧縮文が非文でないか，原文の意味を変えないか．
2. クエリ，他の対して適切に圧縮できているか (情報を落としてないか)．

第 1 段階のエラーに関しては,「文法」,「内容」,「断片」という観点，第 2 段階のエラーに関しては,「過圧縮」,「不要」,「無関係」という観点で分析している．文圧縮という技術は単独の文に対し

て適用する場合と要約のように複数の文に対して適用する場合とでは評価の観点が異なるが，2段階のエラー分析はこの違いをよくとらえており興味深い．

3.2 情報の網羅性に関して

平尾は談話依存構造木の刈り込みによる要約システムが抜粋による参照要約を再現できるかどうかを検証し，テキストの一貫性という点はさておき，参照要約の再現という観点からは談話依存構造木が必ずしも有益でないことを指摘している．

この分析はあくまで参照要約を再現できるか否かのみを論じており，要約として人間が受け入れることが可能かどうかという視点に欠けている．

複数文書要約の対象となる文集合にはしばしばしば同じ情報を伝える文が含まれ，冗長である．要約の情報網羅性を向上させるためにはこうした冗長性を抑える必要がある．こうした観点に対し，小倉は, 文の位置に基づく文スコアの和, 単語重みに基つづく文スコアの和, 要約に含まれる異なり単語数を最大化する文の組合せを選択する多目的最適化問題として要約を定式化し, 異なり単語数を最大化する目的関数が冗長性の削減に貢献したことを定性的に分析している．鈴木は, グラフベースの要約システムに MMR を適用することで冗長性を削減しようとしたが, その効果は安定していなかったことを指摘している．

要約の冗長性の削減に関しては，何を単位として冗長とするのか，それをどの程度まで削減すればよいのかという議論はあまりされておらず，今後の議論が待たれる．

3.3 テキストの一貫性に関して

高村らは談話依存構造木を制約として利用した要約システムとそれを除いた要約システムとを比較し，談話依存構造木が一貫性のある要約の生成に役立ったことを指摘している．

西川は，単一の記事が複数の小トピックから構成される際に，人間の要約作成者はできる限り各トピックを要約に含めるように要約を作成していることを指摘し，どのようなトピックが要約に含まれているかを認定する必要性を指摘している．

要約にテキストとしての一貫性が必要であることは言うまでもないが，一貫性があるが情報の網羅性が低い要約と一貫性にとぼしいが情報の網羅性が高い要約のどちらを人間が好むかなど双方の評価の軸のバランスをどうとらえるかが重要な観点ではないだろうか．

3.4 その他

西川は，原文書と参照要約を比較し，文圧縮だけでは参照要約を再現することができず，言い換え，文融合などの技術が必要であることを指摘している．言い換え，文融合技術はまだまだ自然言語処理分野でも困難な課題であるが自動要約の研究者がこうした課題に取り組むべきであることを示唆している．

4 おわりに

本稿では，Project Next 要約タスク参加者のエラー分析結果をまとめた．要約の誤りは明らかに誤りといえるものからそうでないものまで多岐に渡り，分析が難しい．今回の分析結果からは，文圧縮をそれ単独の技術として利用する場合と要約システムに組み込む場合で評価の観点が異なることがわかった．また，一貫性のある要約を生成するためには，要約シ文間の関係を考慮することが必要であることが示唆された．一方，文間の関係はかなり強い制約であり，それを尊重すると情報の網羅性が下がることも示唆された．情報の網羅性とテキストの一貫性は決して独立した評価軸ではないため，今後は双方の観点がどのような関係にあるかを掘り下げるべきであろう．また，今回の分析には参照要約を利用するものとそうでないものがある．評価の際にどこまで参照要約にたよるべきかどうかも考えていくべき課題であろう．

謝辞

中間報告に対し，有益かつ示唆に富むコメントを下さったアドバイザの皆様，奥村学博士，酒井哲也博士，関根聡博士に感謝致します．

多目的遺伝的アルゴリズムを用いた複数文書要約

小倉 由佳里

2014/11/12

1　要約手法と目的

1.1　目的

抽出型の要約では，要約生成を文の組合せ最適化問題と考えることで，要約生成を行う手法が多く提案されている．要約生成では，内容の網羅性，冗長性，一貫性など考慮すべき要因が複数あり，これらを同時に満たすような要約が求められる．しかし，対象となる文書が増えるほど，考えうる組合せの数は膨大になり，与えられた制約の下で最適化問題を解くのに時間がかかるという問題がある．そこで，本手法では，これらの要因を考慮しながら，実用的な時間で要約を出力するために，組合せ最適化に多目的遺伝的アルゴリズム (多目的 GA) を用いた．多目的 GA で得られる解は，必ずしも最適解ではないが，実用的な時間で準最適解が見つかるという利点がある．

1.2　手法

文の組合せ最適化には，多目的 GA の代表的な手法である NSGA-II 手法の流れは右図の通りである．まず初期個体をランダムに生成する．その個体群に対し，交叉，突然変異の操作を行い，それにより生成された各個体の適合度からランク付けが行われ，上位にランクしている個体のみが次世代へ残る．この一連の操作を設定された世代数分だけ行う．適合度は，設定した適合度関数に基づき算出される．要約生成においては，この適合度関数が，トピックの網羅性，冗長性等の要約の評価指標となる．適合度関数については次章で示す．

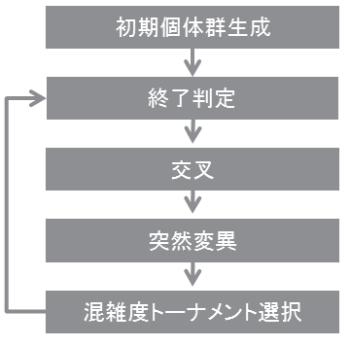


図 1　NSGA-II の手順

また各個体は，文の組合せを表す．図 2 で表すように，個体はバイナリで表現され，各遺伝子座は各文に対応し，遺伝子座の持つ値が “1” の場合はその文が要約に含まれることを示す．

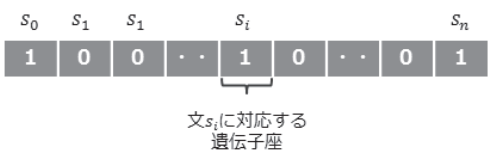


図 2　個体の例

2 目的関数

複数の目的関数を用いて実験を行った．要約に含まれる各文に対するスコアと，文の組合せ全体に対するスコアを用いた (表 2).

	式	
文の位置による文のスコア	$position(S_j) = max\left(\frac{1}{i}, \frac{i}{n-i+1}\right)$ i は文の出現位置， n は文書 D の総文数.	(Lin and Hovy, 1997)
単語の出現頻度による文のスコア	$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{\lvert\{w_i \lvert w_i \in S_j\}\rvert}$ $p(w_i) = \frac{n}{N}$ w_i は文 S_j に含まれる単語， n は単語 w_i の出現頻度，N は総単語数.	(Nenkova et.al., 2006)
要約に含まれる単語の種類	$S = \frac{\lvert\{w_i \lvert w_i \in S\}\rvert}{n}$ w_i は要約 S に含まれる単語， n は要約対象文書の総単語数.	

3 実験と結果

3.1 実験設定

DUC2004task2 のデータセットを用いた.

NSGA のパラメータは，初期個体数 50，世代数 50，交叉率 1.0，突然変異率 0.005 とした.

3.2 実験結果

先に示した目的関数を用いて行った実験では，平均の ROUGE-1 値は 35.85 となった.

文の位置，単語の出現頻度に関するスコア付けのみの場合，全体的に生成される要約が冗長になる印象があった．また，照応解析や，文の順序に関する操作はしていないため，それらに関しては今後の課題となっている.

3.3 出力された要約の例

d30001t

Worried that party colleagues still face arrest for their politics, opposition leader Sam Rainsy sought further clarification Friday of security guarantees promised by strongman Hun Sen. Sam Rainsy wrote in a letter to King Norodom Sihanouk that he was eager to attend the first session of the new National Assembly on Nov. 25, but complained that Hun Sen's assurances were not strong enough to ease concerns his party members may be arrested upon their return to Cambodia.King Norodom Sihanouk on Tuesday praised agreements by Cambodia's top two political parties _ previously bitter rivals _ to form a coalition government led by strongman Hun Sen.

d30006t

The league also called former union director Simon Gourdine to testify, but Feerick upheld union objections and prohibited Gourdine from saying whether it was his understanding when he negotiated the old labor agreement in 1995 that players would not be paid if the owners chose to reopen the agreement and impose a lockout. It's not on us.Despite modest encouragement over a new proposal delivered by the players to the owners, the National Basketball Association Tuesday canceled the first two weeks of the regular season, the first time in the league's 51-year history that it will lose games to a labor dispute.And I believe this is.And I believe this is.

d31009t

ANKARA, Turkey (AP) - Prime Minister Mesut Yilmaz on Wednesday faced intense pressure to step down after allegations that he interfered in a privatization contract and helped a businessman linked to a mobster secure loans.Parliament convened Thursday to vote on whether to move toward a no-confidence motion that could bring down the government over an organized crime scandal.The chances for a new, strictly secular government in Turkey faded Wednesday when a potential coalition partner insisted on giving the Islamic party a share of power.Ecevit refused even to consult with the leader of the Virtue Party during his efforts to form a government.

d31031t

Voting mainly on party lines on a question that has become a touchstone in the debate over development and preservation of wilderness, the Senate on Thursday approved a gravel road through remote wildlife habitat in Alaska.“But I have to say, there hasn't been much evidence of that, this time. Struggling to meet their fourth deadline over the federal budget, congressional Republicans and White House officials wrestled Tuesday with their differences over education, the most politically high-stakes element of the budget battle.“The temptation, said L. Ari Fleischer, spokesman for the House Ways and Means Committee, “is to grab the loot and spend it.

参考文献

[1] K. Deb, A. Pratap, and S. Agarwal. 2002. : A fast and elitist multi-objective genetic algorithm: NSGA2. IEEE Transactionon Evolutionary Computation 6.2, pages 149–172.

[2] C.Y. Lin and E. Hovy. 1997. : Identifying topics by position. Proceedings of the fifth conference on Applied natural language processing.

[3] A. Nenkova, L. Vanderwende, K. McKeown. 2006. : A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 573–580.

成果報告

お茶の水女子大学　小林研究室
修士 2 年　鈴木聡子

研究内容【グラフを用いた時系列文書の要約】

目的

我々が目にする文書は、時間の変遷とともに情報の内容も変化する。例えば新聞記事に着目すると、“エボラ出血熱”などのような 1 つのトピックに関して多くの報道がされている。そのような話題に対して、どのような経緯で話題が変化してきたかを知りたい。本研究では、新聞記事のような時系列文書に着目し、時間の経過に伴う話題の変化の把握が可能となるような要約の生成を目的とする。

提案手法

本研究では LexRank [Erkan et al., 2004]をベースとし、文書が追加されるごとにグラフを更新し、要約が必要なタイミングでのグラフの状態をもとに要約を生成する。
グラフのサイズを制限し、新たに文が追加されたとき、前の時間で下位にある文の情報は捨て、再びランキングを行う。

実験設定

・データセット

Tran らによって提供されるデータセットを用いた。実験に用いたものを表に示す。

Topic	ニュース元	文書数	文数
BP Oil	BBC	293	10395
H1N1	Guardian	76	2630
H1N1	Reuters	207	4769
Haiti Earthquake	BBC	296	9642

これらのデータは、実験の際にストップワードを除き、ステミング処理をする。

・評価

評価には ROUGE を用いた。また、評価には最終的な時点での要約結果を用いる。結果の比較として、ランダムに文を抽出したものと、全ての文を対象に LexRank を用いたものを用意した。

実験結果

ROUGE の値を以下の表に示す。評価は、ユニグラムからトライグラムまで行い、それぞれの Recall と F-score を求めた。
提案手法では、各データセットにおいて 1 番値の高いグラフサイズでの結果を表に記載した。また、各デーセットにおいて評価の際に stopwords を含める場合と含めない場合の両方の値を示す。R1 は ROUGE- 1 を示す。
各手法において 1 番結果の高かったものを太字で示す。

【BP Oil BBC / with stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.594	0.117	0.057	0.565	0.169	0.054
LexRank	**0.729**	0.264	0.086	0.540	0.196	0.064
提案手法	0.698	**0.265**	**0.102**	**0.597**	**0.227**	**0.088**

【H1N1 Guardian / with stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.443	0.092	0.019	0.437	0.091	0.019
LexRank	**0.602**	**0.187**	0.055	0.473	**0.147**	0.043
提案手法	0.593	0.172	**0.062**	**0.476**	0.138	**0.050**

【H1N1 Reuters / with stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.468	0.071	0.014	0.358	0.054	0.011
LexRank	0.616	**0.155**	**0.036**	0.350	0.088	0.020
提案手法	**0.652**	0.152	**0.036**	**0.359**	**0.090**	**0.021**

【Haiti Earthquake / with stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.517	0.128	0.037	0.518	0.128	0.037
LexRank	0.666	0.204	0.069	0.497	0.152	0.051
提案手法	**0.691**	**0.233**	**0.088**	**0.531**	**0.179**	**0.068**

【BP Oil BBC / without stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.417	0.082	0.026	0.417	0.082	0.026
LexRank	**0.576**	0.144	0.053	0.430	0.108	0.039
提案手法	0.525	**0.149**	**0.066**	**0.460**	**0.130**	**0.058**

【H1N1 Guardian / without stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.249	0.045	0.010	0.260	0.047	0.010
LexRank	**0.465**	0.127	0.027	**0.369**	0.101	0.021
提案手法	0.453	**0.130**	**0.036**	**0.369**	**0.106**	**0.029**

【H1N1 Reuters / without stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.270	0.037	0.006	0.222	0.030	0.005
LexRank	0.454	**0.083**	**0.016**	0.273	**0.050**	0.009
提案手法	**0.495**	0.078	**0.016**	**0.284**	0.049	**0.010**

【Haiti Earthquake / without stopwords】

	Recall			F-score		
	R1	R2	R3	R1	R2	R3
Random	0.328	0.052	0.023	0.349	0.055	0.024
LexRank	0.468	0.099	0.037	0.352	0.075	0.028
提案手法	**0.507**	**0.117**	**0.056**	**0.395**	**0.091**	**0.044**

今回の実験では、提案手法は LexRank での結果と近い、もしくはそれを上回る結果を示した。これより文数に制限を与えても、それなりの精度を維持できると考えられる。
しかし比較手法が時系列文書に対応した手法でないことや、時間の経過途中における要約の評価ができていないことが、大きな課題であると考えられる。

MMR の導入

各文をランキングした状態から、重複がないように文を抽出するために MMR を拡張したものの導入を行った。導入式を以下に示す。

$$MMR = \arg\max_{s_i \in S \setminus S'} [score(s_i) - \max_{s_j \in S'} sim(s_i, s_j) \times \eta]$$

S : 要約の候補となる集合

S': すでに要約として抽出されている文集合

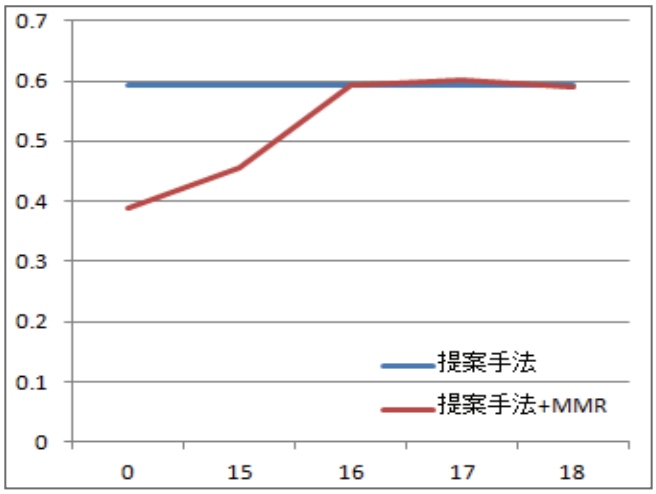
η : 調整係数

要約文の候補の中で、既に抽出されている文に類似している文ほどペナルティが大きくなるようにスコアをつける。２つのデータセットで実験を行った結果を以下に示す。

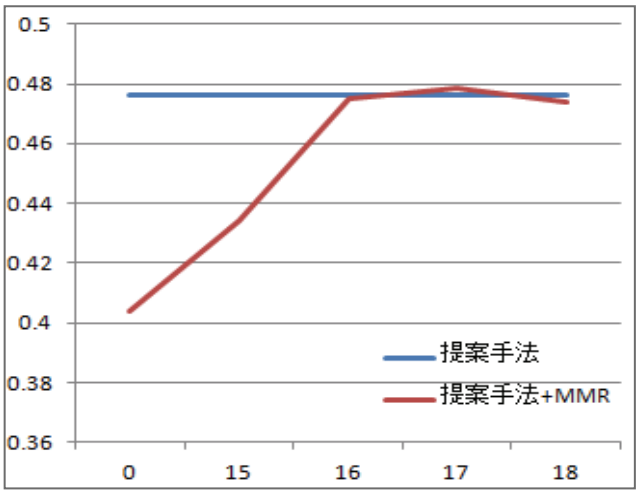
結果

◎H1N1_Guardian(with stopwords)

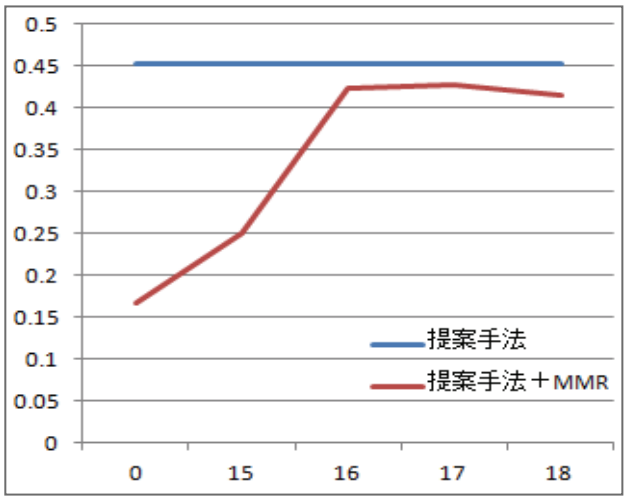
ROUGE-1/Recall

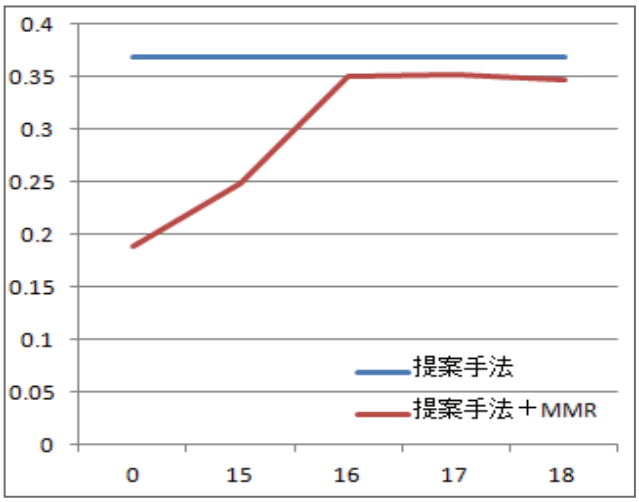


ROUGE-1/F-score

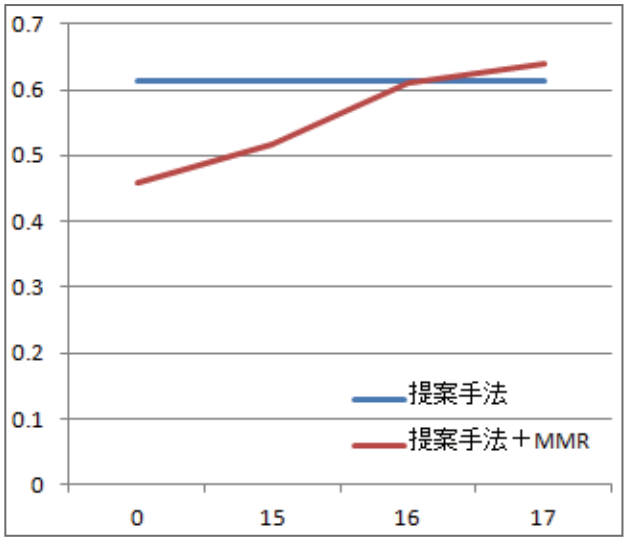


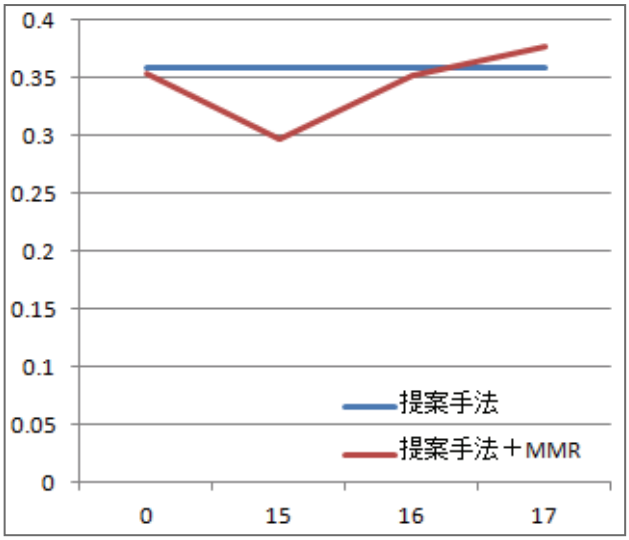
◎H1N1_Guardian(without stopwords)



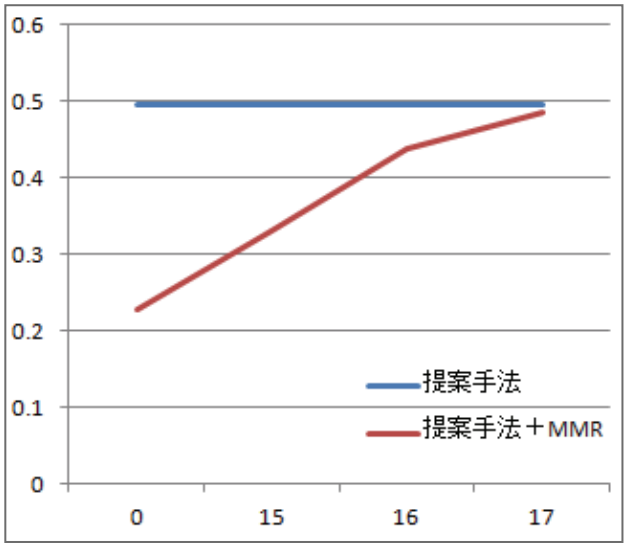


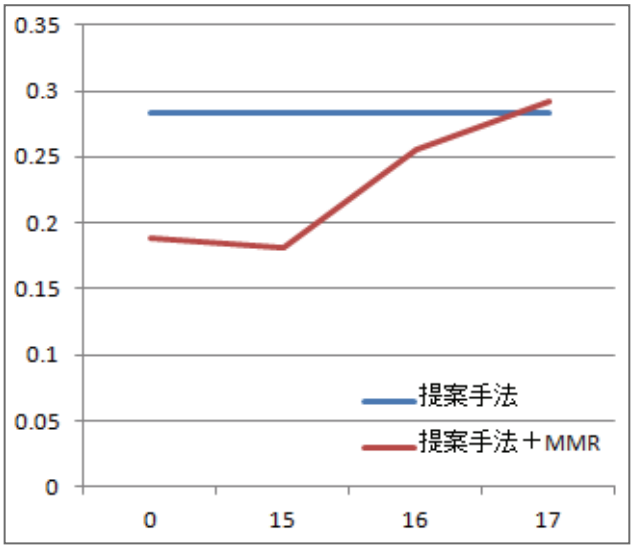
◎H1N1_Reuters(with stopwords)





◎H1N1_Reuters(without stopwords)





青い線は MMR を導入していない提案手法の結果（よって横軸とは関係なく一定値）、赤い線が提案手法に MMR を導入した結果である。

調整係数 η は 10^{-k} となり k がグラフの横軸、ROUGE による評価値が縦軸を表している。

今回、MMR の導入前よりも評価値が上回る部分も所々あったが、多くの場合で同等かそれ以下の結果を示した。
グラフのランキングアルゴリズムでは、中心性により類似した文が上位に集まることが考えられ、ランキング結果からさらに冗長性を考慮することによって、要約結果の改善を期待したが、望ましい結果は得られなかった。

Project Next: 文圧縮を利用したクエリ指向要約を対象にした誤り分析

森田 一
京都大学 情報学研究科

morita@nlp.ist.i.kyoto-u.ac.jp

1 背景

本稿ではエラー分析の対象として，文圧縮を利用したクエリ指向要約を扱う．文圧縮は要約の表現力を高め，原文から不要な情報を除くことができる一方で，誤って文法的でない文や内容の不足した文を生成してしまう．文圧縮の実用性を高めるためには，どのような原因で非文が生成されてしまうのかを分析し，誤って生成してしまう非文を減らすことが重要となる．特にクエリ指向要約では，文の大部分がクエリと関連していない原文から，一部を取り出して要約を作る必要がある場合があり，文圧縮を適切に行うことが重要となる．クエリ指向要約において，文圧縮が文のどこを削除するかは，クエリとの内容的な関連性と文法性を考慮して決定する必要がある．このため，クエリ指向の要約における文圧縮のエラーについて，その両方の観点から分析を行う．

2 設定

本稿では 森田 [1] で提案した手法の出力について分析を行った．評価にはデータセットとして，NTCIR-8 ACLIA2 の 100 クエリのうち factoid 型の質問を除いた 80 クエリと，そのクエリに対する回答ナゲット，および対応する毎日新聞の記事セットを用いた．

3 分析

3.1 分析方法

クエリに対する関連性を含む内容的な側面と文法的な側面の 2 つの観点から分析を行うため，二段階に分けて分析を行う．一段階目の分析では文圧縮単体について分析を行った．ここでは文法的に圧縮ができているかどうか，生成された文の解釈が変えてしまっていないかを評価し，エラーの分類を行った．二段階目では与えられたクエリに対して，また要約中の他の文に対して適切に圧縮出来ているかを分析した．ここでは主に，必要な情報が要約文から抜け落ちていないか，明らかに不要な箇所が含まれていないかを評価した．

3.2 一段階目の分析

一段階目では，出力に含まれた全ての文について，以下の５つに分類を行った．

- 文法的に問題がある (文法)
- 内容的に問題がある (内容)
- 短く圧縮されすぎて意味を読み取ることができない (断片)
- 圧縮されていない (非圧縮)
- 文法的に問題なく圧縮できている (問題なし)

文法的に問題があるとした文には，主に圧縮により文の構造が保たれなくなった例が含まれる．文の構造が保たれなくなった結果，生成された文自体は非文でなくとも，元文と示す意味内容が変わってしまう場合は，この分類 (文法) に含めた．以下，文圧縮により削除された箇所を斜線で示す．次の文では，“中間線から” が係るべき “数キロ中国側の” が除かれてしまったために，文の構造が変わってしまっている．

> 東シナ海のガス田開発は、中国が03年8月、日中中間線から数キロ中国側の「春暁」で着手。

(内容) として分類したのは文の構造は保たれているものの，必要な修飾を除いてしまったことにより，文が不自然であったり解釈が困難になっている文である．文法的に問題がある場合と同様に，元の文と解釈が異

なってしまうものも含まれる．照応先が圧縮により除かれてしまっている場合は内容が不明確になるが，要約中の前の文やクエリにより解釈が可能であるなど単文では評価が難しいため今回は対象外とした．次の例では“米国と”と“日本、”の部分が削除されてしまったことで，６カ国が指す国から日本が抜け，旧ソ連が入っているかのような文になってしまっており，元の文と解釈が異なってしまっている．

> ８５年に米国と旧ソ連が合意し、日本、米国、韓国、中国、ロシア、ＥＵの６カ国・地域が協力して１基を建設する国際プロジェクトとなり、青森県六ケ所村と仏カダラッシュが候補地だった。

(断片) と分類したのは，次の例のように，ほぼ単語しか残らない圧縮がなされてしまっている場合である．文法的な問題と内容的な問題の両面を併せ持っており，ほとんどの場合には元の文の意図を解釈することはできなくなっている．

> ハンセン病の歴史に詳しい藤野豊・富山国際大助教授（日本近現代史）の 話

次の文はプーチン大統領がどのような人物か，というクエリに対して要約を生成しており，クエリの求めている情報を出力している．このように，まれに１単語でもクエリに対する応答として十分である場合もあるが，例外的であるためここではクエリに対して適切かどうかは考えず短く圧縮され過ぎていると分類した．

> イ・ス・ラ・エ・ル先行 高支持率

(非圧縮) は要約の際に文全てが使われ文圧縮されなかった文，(問題なし) は適切に文圧縮が行われた文を表す．以下の表 1 に各分類に含まれた文数を示す．

表 1: 文圧縮のエラータイプ

(文法)	13 文
(内容)	14 文
(断片)	36 文
(問題なし)	54 文
(非圧縮)	84 文
合計	201 文

(文法) に属するエラーに分類した 13 文のうち 9 件は係り受け解析に失敗しており，文圧縮でその係り受けの保存に失敗していることが主な原因となっている．係り受け解析の失敗以外の要因としては，次の例のように文圧縮する際に書き換えが必要となる場合が存在した．

> 写真は草刈郁夫、近藤卓写す。格解析結果:
> 写真ヲ 郁夫ガ 卓ガ 写す

この例では，“草刈郁夫、近藤卓”を取り除く場合には，“写真を”と書き換えなくては意味が変わってしまう．しかし，今回用いたモデルでは書き換えを行うことができないため，文圧縮を行った際に元と文の意味が変わってしまう要因となった．他の要因としては，括弧で囲まれた名詞が前の名詞句を修飾していたが，括弧内の名詞だけを取り出してしまった場合や，“〜による”のような機能的な複合辞に対しては格解析を参照しようとするのではなく，“による”に係る語を必須とするべきであったが，その処理が不足していた場合などがあった．

(断片) に属するエラーは，要約を生成するアルゴリズムと，問題設定あるいは評価方法に起因する問題である．現在の要約の評価では一定の要約長の中にできるかぎり多くの語を含めることが評価指標上はプラスとなるため, 生成アルゴリズムは要約長が残り数文字となった場合にも，その数文字に収められるように無理のある文圧縮を行う．読みやすさの観点からは，残り文字数が一定以下となった時点で生成を打ち切るなどの対応が必要となる．

(内容) に属するエラーは，圧縮の際に必須格として保護できずに削除してしまっている例が 14 例中 4 例，括弧の対応を崩してしまう例が 2 例，形態素解析のエラーにより人名が分割され名前の一部のみを取り出してしまう例が 1 例 (“アッバス首相” > “バス首相”), 固有名詞を分割してしまった例 (“千と千尋” > “千尋”) が 2 例あった．残りの例では，形式名詞等に対する連体修飾を除いてしまったために，不自然な文になっていた．

3.3 二段階目の分析

二段階目の分析はより詳細に分析を行うため，(内容) の 14 文と (問題なし) のうちからランダムに同数の 14 文を選んで分析を行った．二段階目では，与えられたクエリに対して文から不要な箇所を除き適切に圧縮できているかどうか，情報の過不足のみを問題とし，分析を行った．ただし，クエリが求めている情報が原文に含まれていても，要約中の他の文で与えられ

ている情報は，不要な情報として扱う．二段階目では情報の過不足について，以下の 4 つに分類した．

- 必要な情報を削除している (過圧縮)
- 不要な箇所を含む (不要)
- 原文に必要な情報がもともとない (無関係)
- 適切に圧縮できている (適切)

以下の表 2 に各分類に含まれた文数を示す．一文が (過圧縮) と (不要) の両方に当てはまる例は (内容) の 14 文の内に 3 文あり, 集計では (過圧縮) と (不要) の両方に 1 文として重複してカウントした．

表 2: 文圧縮のエラータイプ

エラータイプ	(内容)	(問題なし)
(過圧縮)	7 文	1 文
(不要)	5 文	6 文
(無関係)	5 文	3 文
(適切)	0 文	4 文

(過圧縮) は (内容) で述べた問題の他に，(問題なし) とされた文圧縮でもクエリとの関連性がうまくスコアに反映されていないために必要な部分を除外してしまう場合があった．このため，クエリとの関連性スコアをより正しく評価できるよう，クエリとの関連性スコアをより改善していく必要がある．

(不要) では，文圧縮の手法が係り受け木のルートを必ず必要とするため，モデル上圧縮できない例が 11 文中 7 文あった．これらの例を正しく文圧縮するためには，係り受け木中の任意の部分木を出力できるモデルへ拡張する必要がある．

(無関係) の場合には，クエリと関連しているかどうかという情報を文圧縮に用いることができないため，(内容) に分類されるエラーが多いことを想定していたが，実際にはやや多い程度で大きな差はなかった．(内容) の問題を解決するためには，クエリとの関連がある場所を残すだけではなく，必須格の判定をより精度高く行うことや，固有名詞等の分割してはいけないものを認識して文圧縮に反映することが必要となることが分かった．

4 まとめ

生成された要約で，文圧縮の失敗など明確にエラーと呼べるものは，基礎的な解析の誤りであったり，解析結果を適切に利用出来ていないことが原因であった．基礎的な解析を積み重ねることが重要であることが分かった一方で，今回の分析は表層的な問題の原因を基礎的解析に対応付けただけであるとも言える．今後自動要約をより良いものにするためには，主観的な分析とともに，何らかの応用タスクを定めて分析を行うことも必要となるであろう．

参考文献

[1] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Subtree extractive summarization via submodular maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1023–1032, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

単一文書要約における談話構造の効果と課題

高村大也 † 菊池悠太 ‡

1 はじめに

本稿では，東工大チームが行った文書要約の誤り分析について報告する．特に，文書要約において談話構造がどのような効果を持つか，また談話構造を用いた場合の課題は何か，などについて焦点を当てる．

2 実験設定

我々は，Kikuchi ら [2] の要約手法による要約結果に対する誤り分析を行った．Kikuchi らの手法は，修辞構造理論 (Rhetorical Structure Theory; RST)[4] による文間関係により文ノードを結合することで構築できる談話木において，各文ノードにその文の構文木を対応させ，全体を入れ子木として表現した上で，この入れ子木からの根付き部分木抽出問題として，要約を定式化したものである．ただしここでは，要約内の各文の文法性ではなく，要約の文章としての適切さに分析の焦点をあてるため，文短縮は行わないことにした．すなわち，談話木の根付き部分木抽出問題を解くことで，文抽出による要約生成を行う．Kikuchi らの手法を文選択にしたものに加え，そこから RST の情報を除いた手法を試した．RST の情報は，最適化問題において，談話木の根付き部分木を実行可能解とする制約として用いられている．よって上述した二つの手法のうち後者は，その制約を除去した最適化問題として定式化されることになる．

データは，Kikuchi ら [2] が使用した RST Discourse Treebank (RST-DTB)[1] の一部を用いた．実際に分析したのは 16 文書である．これらの 16 文書は，RST を用いた場合の ROUGE 値 [3] が低いものである．ただし，文書 1143 については，制約を満たす解が存在しなかったため，分析から外している．これは，談話木の根に対応する文は必ず選択されるが，この文が非常に長く，その一文だけで要約長制約に違反してしまうような事例であった．要約長は参照要約長以下となるように制約を与えた．

†東京工業大学，takamura@pi.titech.ac.jp
‡東京工業大学，kikuchi@lr.pi.titech.ac.jp

3 誤り分析の枠組み

我々の分析では，要約の誤りは四種類に大別され，一部の種類はさらに細分化される：

- **非文**：文単位の文法誤りが含まれる，
- **非適格文章**：文章として不自然な箇所がある，
  - **論理構造不明**：文と文の論理的なつながりが不明な箇所がある，
  - **照応解決困難**：照応の解決が困難な箇所がある，
- **文意の変化**：元文書と異なる意味で解釈できる，
  - **論理構造変化**：文と文の論理的なつながりが変化してしまう箇所がある，
  - **照応解決誤り**：照応が誤って解決されてしまう箇所がある，
- **重要内容の同定誤り**．

我々の誤り分析は，大枠では西川の誤り分類 [5] に従っている．異なる点は，西川の分類では非文と非適格文章が一つにまとまっていることと，我々は，非適格文章および文意の変化のそれぞれを，論理構造の問題に起因するものと，照応の問題に起因するものに細分化していることである．

今回の誤り分析は，文選択により生成された要約に対して行うので，非文となることは少ないが，文分割に失敗することで非文となる場合がある．

4 分析結果

分析結果は表 1 の通りである．各要約には複数の誤りラベルが付きうるので，表の各行の和は異なりうる．

また，実際の要約例などは付録に記載する．全節の誤りの種類に加え，最も重要な内容は同定できているが，対応する文が過度に長いために他の重要な内容を含めることができなかった例もあり，これについては付録ではその他として記載している．また，うまく

表 1: 分析結果

手法	非文	非適格文章		文意の変化		重要内容の同定誤り
		論理構造	照応	論理構造	照応	
RST なし	1	12	4	1	0	12
RST あり	1	1	3	0	0	10

いかなかった要約課題を分析対象とするため，自動評価指標である ROUGE 値が低い要約課題を選んでいるが，実際は良い要約が生成できているにも関わらず ROUGE 値が低くなってしまっている場合もある．これについても付録ではその他として記載している．

まず，論理構造不明な非適格文章の数が，RST ありの場合は 1，RST なしの場合は 12 と大きく異なっている．これは，RST から得られる談話構造を利用している Kikuchi らの手法が，有効に働いていることを意味している．RST を用いた場合の要約の成功例としては，文書 1128 を挙げておく．RST を用いているにも関わらず論理構造が不明となっている文書 1121 については，RST 情報に誤りがあると思われる．

また，RST の有無に関わらず，照応解決は問題になっている．例えば，文書 1302 では “such” が何を指しているのかが不明である．要約生成において適切な参照表現を生成することは，重要な研究課題である．

今回用いたデータにおいては，文意が変化するようなケースは非常に少なかった．唯一の例は，RST なしの場合の文書 2317 であった．文書 2317 では，元文書において副詞 “also” を含む文があり，それが直前の文と共通点を持つことを “also” が示していた．副詞 “also” を含む文は要約に選択されていたが，要約を読むとこの文が別の文と共通点を持っているかのように解釈されてしまい，文意が変化していると判断した．

また，重要内容の同定誤りについては，RST の有無に関わらず多数あった．RST ありの場合については 10 文書のうち 3 文書が，RST なしの場合については 12 文書のうち 3 文書が，数値情報が原因となっていた (例えば文書 1154)．これは，実際の数値情報が細かく述べられている文が選択されているような場合である．これらに対応する参照要約では，数値情報は細かく述べられておらず，その数値に対する解釈やそのような数値に至った背景などが述べられている．数値情報に対してその解釈や背景を抽出あるいは生成する技術の開発が，今後の研究課題として挙げられる．また，RST ありの場合は，談話木の根である単文が長いことにより，他の重要内容が十分に被覆できていない事例が 3 つあった．これは文圧縮の重要性を示唆している．これら以外の文書については，もう少し深い意味解析が必要となるだろう．

また，非文が含まれている場合が 1 つあったが (文書 1154)，これは前処理の文分割誤りによる．

5 おわりに

文書要約の誤り分析について報告した．特に，文書要約において談話構造がどのような効果を持つか，また談話構造を用いた場合の課題は何か，などについて焦点を当てた．誤り分析の結果，ROUGE のような単語レベルの自動評価指標では十分に良さが測れないものの，談話構造は要約生成に非常に有効であることがわかった．また，照応解決を要約に適切に組み込むこと，また数値情報を解釈しその解釈内容を生成する技術を開発すること，などが次に取り組むべき課題として重要であることがわかった．また，ROUGE の評価指標としての限界を示唆する結果も得られており，新たな評価指標の開発も今後の課題として重要である．

参考文献

[1] Carlson, L., Marcu, D., and Okurowski, M. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGdial*, pp. 1–10, 2001.

[2] Kikuchi, Y., Hirao, T., Takamura, H., Okumura, M., and Nagata, M. Single document summarization based on nested tree structure. In *ACL 2014, Short Papers*, pp. 315–320, 2014.

[3] Lin, C. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

[4] Mann, W., and Thompson, S. Rhetorical structure theory: toward a functional theory of text organisation. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[5] 西川仁. 自動要約タスクにおける誤り分析の枠組みの提案. ProjectNext 要約タスク, 2015.

付録

元文書と，それぞれの手法での要約結果の例を示す．また，各要約の ROUGE 値や，我々の分析結果は要約のあとの [] 内に記載する．文書は，RST を用いた手法の ROUGE 値に関し，昇順に並べる．

ID	参照要約	RST なし	RST あり
1302	Lloyd's of London, the long time pillar of the insurance industry, is facing difficult times. The company has been hit by huge claims from a ten-year string of disasters and is also losing investors at an unprecedented rate. The company is slowly making some changes but it must face the challenge of changing some of its long-standing traditions, especially the failure to embrace modern technology,	In 1985, it was 2.1% . Between 1981 and 1986 , the most recent five-year period for which figures are available, Lloyd's reported over # 3.6 billion in claims and reserves against future losses ($ 5.7 billion at today's exchange rates), more than double the # 1.35 billion posted in the previous five-year period. [**11.4%, 重要内容の同定誤り 非適格文章 (論理, 照応)**]	At Lloyd's of London, underwriters still scratch out policies using fountain pens and blotting paper. Such trappings suggest a glorious past but give no hint of a troubled present. Unease is widespread among exchange members. Lloyd's has endured decades of genteel decline. Even Lloyd's specialty – unusually risky ventures – is being challenged. Lloyd's is moving forward on some fronts, though. [**8.6%, 非適格文章 (照応) その他 (評価)**]
1154	Franco Zeffirelli's production of La Traviata at the Metropolitan Opera House featured lackluster performances by singers Edita Gruberova and Neil Shicoff. There were also numerous problems with the set. The Met has been able to count on Zeffirelli's productions regardless of the cast. The New York City Opera, by contrast, has had to cancel its season due to a labor dispute.	though but it made its mark: Says director George Miller ("Mad Max"): "I think the Tokyo festival may become known as a major attraction for young directors because of the money as well as the recognition." He added: "Banks must open their books." "The price can not be less than $7,000. " [**0.0%, 重要内容の同定誤り 非文 非適格文章 (論理)**]	Cast as Violetta Valery in a new production of Verdi's "La Traviata, "Ms. Gruberova last week did many things nicely and others not so well. though but it made its mark: Since Mexican President Carlos Salinas de Gortari took office last December, special agents have arrested more than 6,000 federal employees on charges ranging from extortion to tax evasion. [**15.2%, その他 (前処理)**]
2317	Elcotel Incorporated expects lower second quarter earnings as compared to 1988 but earnings should improve in the second half of the fiscal year. Company Chairman and CEO George Pierce said the earnings drop reflects a twelve month decline in the sale of privately owned pay phones, Elcotel's main business. Earnings should improve as the company moves into other areas, most notably, public facsimile machines. Two other new areas will be automatic call processing and convenience telephones.	Revenue totaled $5 million. Mr. Pierce said Elcotel should realize a minimum of $10 of recurring net earnings for each machine each month. Elcotel has also developed an automatic call processor that will make further use of the company's system for automating and handling credit-card calls and collect calls. Mr. Pierce estimated that the processors and convenience phones would produce about $5 of recurring net earnings for each machine each month. [**25.5%, 文意変化 非適格文章 (論理)**]	Elcotel, a telecommunications company, had net income of $272,000, or five cents a share, in its year-earlier second quarter, ended Sept. 30. Revenue totaled $5 million. George Pierce, chairman and chief executive officer, said in an interview that earnings in the most recent quarter will be about two cents a share on revenue of just under $4 million [**17.0%, その他 (長さ調整失敗)**]
1121	The De Beers Company diamond dig, located in the Namibian desert, is one of the world's most lucrative. Operations last year produced 934,242 carats, most of them gem quality. The restricted mining area is one of the most desolate locations in Africa but the mine headquarters at Oranjemund boasts numerous amenities.	Men would crawl in the sand looking for shiny stones. It was as easy as collecting sea shells at Malibu. Oh yes, the Atlantic was also pushed back 300 yards. And Oranjemund boasts attractions besides diamonds. Mechanized vacuum cleaners probe the sand like giant anteaters; most of the diamonds are still found in the sand [**9.7%, 重要内容の同定誤り 非適格文章 (論理)**]	But only after a fleet of 336 gargantuan earthmoving vehicles belonging to De Beers Consolidated Mines Ltd., the world's diamond kingpins, do their work. Still, miners from all parts of Namibia as well as professional staff from De Beers's head offices in South Africa and London keep coming. [**19.4%, その他 (前処理)**]
1128	The Fuji apple may one day replace the Red Delicious as the number one U.S. apple. Since the Red Delicious has been over-planted and prices have dropped to new lows, the apple industry seems ready for change. Along with growers, supermarkets are also trying different varieties of apples. Although the Fuji is smaller and not as perfectly shaped as the Red Delicious, it is much sweeter, less mealy and has a longer shelf life.	"The Fuji is going to be No.1 to replace the Red Delicious," he says. New apple trees grow slowly, and the Red Delicious is almost as entrenched as mom. A good Delicious can indeed be delicious. More than twice as many Red Delicious apples are grown as the Golden variety, America's No.2 apple. "I've got 70 kinds of apples. "There's a Fuji apple cult. [**42.4%, 重要内容の同定誤り 非適格文章 (論理)**]	A Japanese apple called the Fuji is cropping up in orchards the way Hondas did on U.S. roads. Some fruit visionaries say the Fuji could someday tumble the Red Delicious from the top of America's apple heap. It has a long shelf life and "doesn't fool the public," says Grady Auvil, an Orondo, Wash., grower who is planting Fujis and spreading the good word about them. [**30.3%, その他 (評価)**]

Project Next Summarization : 自動要約タスクにおける誤り分析の枠組みの提案

西川 仁
NTT メディアインテリジェンス研究所
nishikawa.hitoshi@lab.ntt.co.jp

1 はじめに

本稿では自動要約の誤り分析を扱う．

自動要約研究の題材として広く用いられるコーパスは概ね，数十から数百の，入力文書と参照要約の組からなる．自動要約の入出力はいずれも複数の文からなる文章であり，機械翻訳のように文ではなく，また自然言語解析のように何らかの中間表現でもない．そのため，誤りの分析において考慮しなければならない要素が多く，数十といったオーダでも，必ずしも ROUGE [3] などの定量的な評価指標による表層的な分析以上の分析が，十分になされているとはいいがたいというのが筆者の個人的な印象である．そのため，何らかの誤りを含むと思われる要約をどのように分析し，体系的な方法論は存在せず，したがって自動要約分野の研究者が各々の方法論をもって分析を行っているのが現状と思われる．

本稿では，自動要約における誤り分析の枠組みを提案する．まず，要約器が作成する要約が満たすべき 3 つの要件を提案する．また，要約器がこれらの要件を満たせない理由となる 5 つの原因を提案する．3 つの要件と 5 つの原因から，15 種類の具体的な誤りが定義され，自動要約における誤りはこれらのいずれかに分類される．

本稿の構成は以下の通りである．2 節では自動要約研究における誤りについて簡単に述べ，本稿で議論を行う範囲を明らかにする．3 節では自動要約における評価指標をいくつか紹介し，誤りを検出する方法を述べる．4 節では誤り分析の枠組みを提案し，検出された誤りが提案する 15 種類のいずれかに分類できることを示す．5 節では実際の要約例を分析し，要約に含まれる誤りを提案した枠組みに基づいて分類した結果を示す．6 節では本稿をまとめ，今後の展望について述べる．

2 自動要約の誤り

一般に，誤りといえば，本来得られるべき何らかの正しい結果があるものの，それとは異なる，すなわち正しくない別の結果が得られた際にそれを指していうものと思われる．文書分類であれば与えられた文書を正しい分類先に分類できなかった際にそれを誤りということができる．そのため，何らかの正しい結果，すなわち正解が定まらなければ誤りも定めることができない．

自動要約においては，この正解（以下，参照要約と呼ぶ）をいささか一意に定めづらい．これは，自動要約に限らず，自然言語生成を目標とする研究に共通する問題であるが，自動要約課題において，複数の人間の作業者に参照要約の作成を依頼すると，作業者に与える指示にもよるものの，まったく同一の参照要約が作成されるということはまずない．そのため，ある参照要約を基準とした際には誤りとなる要約が，別の要約を基準とした際には誤りとならないことがある．

本稿では，この問題は脇に置く．すなわち，ある 1 つの参照要約が存在するとき，それと要約器が作成した要約（以下，便宜的にこれを機械要約と呼ぶ）を比較し，その差分を誤りとする．すなわち，何か差分があれば誤りを含むし，そうでなければ誤りを含まない．誤りについては次節にて述べる．この単純化は以下の理由に基づく：

- 単一の参照要約の誤り分析の枠組みが存在しない状況において，複数の参照要約が誤り設定の枠組みを設定するのは困難であると考えられること．
- 単一の参照要約の誤り分析の枠組みを設定できれば，それに基づいて複数の参照要約が存在する場合を検討することができると考えられること．

これらの点から，本稿でのこの単純化は，問題の過度な単純化ではなく，合理的な問題の分割であると考

える．

自動要約の対象となるある入力文書とその参照要約，さらに要約器が作成した機械要約の 3 つ組が存在するとき，それらを入力とする何らかの評価関数が存在すれば要約器が作成した要約が誤りを含んでいるか否かを判定できる．次節ではこのような評価関数について述べる．

3　自動要約の評価

参照要約と機械要約がそれぞれ与えられた際に，それを比較することで要約に誤りがどの程度含まれるか評価することができる．自動要約の評価の観点は大きく内容性と可読性の 2 つにわかれ，したがって評価尺度も大きくわけてそれらのどちらか一方を評価するものとなっている．本節では 4 種類の評価尺度を取り上げる： Precision/Recall [9]， ROUGE [3]， Pyramid [6]， DUC Quality Questions [5] である．前の 3 つは主に内容性を評価する尺度であり，最後の 1 つは可読性を評価する尺度である．

3.1　Precision/Recall

Precision（適合率）と Recall（再現率）は自動要約課題を重要文抽出課題として捉えた際の第一の評価尺度である [9]．人間の作業者が重要文として選択した文を機械が選択した割合は適合率と再現率の観点から求めることができる．この尺度に基づいて，入力文書が含む重要な情報を機械が認識する性能を評価することができる．

3.2　ROUGE

ROUGE は参照要約と機械要約の n-gram の頻度分布を比較する [3]．ROUGE は単純ではあるが，人間の作業者による手動の評価と強い相関を持っており [3]，自動要約の評価において広く用いられている．

ROUGE の問題は， n-gram の頻度分布が参照要約とまったく同一となるワード・サラダを生成した場合，そのワード・サラダが参照要約とまったく同一のものとして評価されるという点にあり，可読性の評価は必ずしも得意ではない．

3.3　Pyramid

Pyramid 法は，人手によって複数の参照要約に含まれる等価な情報 [1] を同定し，多くの参照要約において出現している情報の大きい重みを与え，大きい重みを得ている情報を多数保持する機械要約に高い評価を与える [6]．

Pyramid 方は人手によって行われるため高コストであるが，その分，正確な評価が可能である．

3.4　DUC Quality Questions

DUC Quality Questions は要約の Readability（以降，本稿では可読性と呼ぶ）を評価する尺度である．DUC Quality Questions は自動評価のためのものではなく，人間が要約の可読性を評価する際に利用される一種の手引きである．要約の可読性が劣悪である場合は，これらの項目のいずれかあるいは多くにおいて低い得点を得ている可能性が高い．

これらの評価尺度によって，機械要約が誤りを含んでいることを明らかにできれば，次に行うべきことはどのような誤りが中に含まれているのか調査することである．次節はこの調査について述べる．

4　誤り分析の枠組み

何らかの入力文書から要約器が作成した要約が誤りを含んでいるということがわかれば，それは誤り分析の対象となる．

4.1　自動要約の誤りの種類

これまで自動要約研究に携わってきた経験から，筆者は以下の 3 つの原則を自動要約器は満たすべきと考える：

1. 出力から情報を読みとれること．情報を読み取れないような文が出力されていないこと．この点は自動要約の可読性評価と概ね対応する．情報を読み取れないような文が出力された場合には，以下の 3 つのケースが考えられる．

[1] 情報の単位は様々であるが，その単位は最大でも節以下である．

(a) 要約がユーザの要求とは異なる言語で出力されている場合や，要約器がその内部処理において利用している制御記号などが出力されており，要約から文意を読み取れない場合．何らかの理由により要約が出力されない場合も含む．

(b) 文法的でない文（非文）が要約を構成しており，要約の文意が取れない場合．

(c) 個別の文は文法的であるが，要約を構成する文同士の論理関係などが明らかでなく，全体として文意が取れない文章が要約となっている場合．

本稿ではこれら 3 点をまとめて，便宜的に「非文章」と呼ぶ．

2. 出力から読み取れる情報が，入力および読み手の希望を鑑みて，重要であると思われること．重要でない，枝葉末節の情報が出力に含まれないこと．この点は自動要約の内容性評価と概ね対応する．

3. 読み取れる情報が，入力と矛盾せず，入力が出力を含意すること．読み手が入力を読んだ際と出力を読んだ際に異なる結論に至らないこと．

これらの原則から，自動要約の誤りの分析において 3 つの観点が導出できる：

1. 要約器が出力した文章から文意が読み取れるか．

2. 出力から読み取れる情報が，入力および読み手の希望を鑑みて，重要であるか．

3. 読み取れる情報が，入力と矛盾せず，入力が出力を含意するか．

この 3 つの観点が，要約器の誤りを考える際に，最初の分類としてあらわれるものと思われる．

4.2 要約器の誤りの原因

一方，要約器が前節のように誤る原因には以下の観点が考えられる：

1. 操作の不足：要約器が，人間の作業者がテキストに対して施す操作と同等の機構を保持してないことに伴って生じる誤り．言い換えなどの操作ができないために入力された文を短縮することができず，人間と同等の情報量を要約に含めることができない場合や，要約器が入力された文において省略されているゼロ代名詞を復元できず要約の文意を損なう場合が含まれる．

2. 特徴量の不足：特徴量が不足している場合．この場合は 2 つにわけることができる．

   (a) 特徴量の設定不足：要約器において設定されていない特徴量が要約の作成において重要な役割を果たすと思われる場合．段落に関する情報を入力されたテキストから得ることができるのにもかかわらず，要約器はそれを特徴量として認識していない場合など．

   (b) 言語解析の失敗：解析器が誤り，特徴量として設定されている情報が正しく取得できなかった場合．固有表現認識器が固有表現を取り損ね，要約器がそれを特徴量として利用できなかった場合など．

3. パラメタの誤り：訓練事例の不足，不適切な学習手法の利用などによって推定されたパラメタが十分でない場合．

4. 探索の誤り：正しいパラメタが得られているが，探索誤りを生じ誤った要約を生成した場合．本来はより良好な文の組み合わせがあるにもかかわらず，探索誤りによって不適切な文の結果を出力として選択した場合など．

5. 情報の不足：そもそも要約器に対して入力された情報だけでは参照要約まで到達できない場合．人間の要約作成者が入力以外の情報源を利用して要約を作成した場合など．

以上の，3 種類の誤りの種類と，5 種類の誤りの原因から，自動要約における誤りは 15 種類のいずれかに分類できると期待できる．これをまとめたものを表 1 に示す．

なお，これらとは別に，そもそも参照要約が信頼できないと思われる場合（参照要約作成者の読みが誤っていると思われる場合など）もありうるが，ここではそれは除外し，あくまで参照要約が正しく，機械はとにかくそれを模倣することを考えればよいという場合を想定した．

5 分析の実践

本節では前節で提示した分析の枠組みを，本稿で分析の対象とした文書に対して適用する．

まず，分析の枠組みの適用の対象とする機械要約を作成する．次に，それらに対して人手による分析を行い，その後分析の結果を提案した分析の枠組みに基づいて整理する．

分析に際しては，特に佐久間らによる人間の要約の分析 [8] を参考にした．また，Luhn [4] や Edmundson [2] による初期の分析も同様に参考にした．

5.1 実験設定

5.1.1 データ

TSC-2[2] のフォーマル・ランのデータを用いた．その中でも作成者 1 による自由記述の要約を参照要約として取り上げ，特に，文書番号 990305053 を用いた．

5.1.2 要約器

要約器については，西川らによる単一文書要約器 [7] を利用した．文短縮は用いずに利用した．

5.2 結果

表 3 に入力文書（文書番号 990305053 ）を示す．太字は入力文書と参照要約とで文アライメントを取り，対応づけが取れた文同士において共通の単語である．下線は要約器によって重要文と認定された文である．表 4 に参照要約を示す．分析の対象となると思われる点について下線を加え，どのような現象が生じているか下線で示された部分の後に加筆した．表 5 に自動要約を示す．太字は自動要約と参照要約とで文アライメントを取り，対応づけが取れた文同士において共通の単語である．表 4 と同様に分析の対象となると思われる点について下線を加え，どのような現象が生じているか下線で示された部分の後に加筆した．表 2 に入力文書および参照要約，自動要約の統計量を示しておく．

表 2: 入力文書および参照要約，自動要約の統計量

	文数	文字数
入力文書	33	1215
参照要約	15	495
自動要約	11	493

[2]http://lr-www.pi.titech.ac.jp/tsc/tsc2.html

5.3 誤り分析

5.3.1 重要部の同定の失敗

まず， ROUGE-1 [3] の 値は 0.385 であった．文単位でみると，自動要約に含まれる文のうち，完全に参照要約に含まれない文は 2 文めと 11 文のみであり，11 文中 2 文にとどまっている．このことから，要約器の精度（適合率）は $\frac{9}{11}$ に達しており，要約器は高精度に重要文を同定していることがわかる．一方，再現率の観点から見ると，参照要約は入力文書 33 文のうち 15 文を要約として採用しており [3]，再現率は $\frac{9}{15}$ に留まっている．このことから，単語単位での再現率の指標である ROUGE-1 の値は，まだまだ改善の余地があることがわかる．

次に，重要部同定の失敗の原因を探る．表 3 を見ると，要約器は特に後半の文を選択できていない．入力文書において，どのよう話題が遷移しているかを表 6 に示す．

表 6: 入力文書に含まれる話題の遷移

話題番号	文	話題
1	1-2	全人代の開催
2	3-4	朱首相の中国の改革に対する決意
3	5-10	中国の改革に対する熱気の薄れ
4	11-16	中国に対する信頼を揺らぎ
5	12-16	統計値の水増しの疑い
6	17-21	金融改革における外資の取り扱い
7	22-31	香港に対する官僚的な対応
8	32	記事のまとめ

全人代が開催されるということ（話題 1）と中国の改革とその行く末が危ぶまれるということ（話題 2-4）と，その具体的な例（話題 5-7）が並んでいる．参照要約を見ると，参照要約の作成者はできる限りこれらの情報を網羅的に要約に含めることを狙っていることが読み取れる．要約器が後半の文を選択できなかったのはこのような話題の構造を理解することができなかったためで，この構造を要約器に理解させることは重要部の同定に決定的に重要である [4]．

[3]2 つの文を 1 つの文としてまとめているケースがあり，そのため参照要約は 13 文から構成されている．詳しくは文融合の節にて詳述．

[4]なお，西川らの要約器ではこのような話題の遷移を段落を通じて獲得しているが，毎日新聞データでは段落に関する情報が失われ

5.3.2 文の融合

異なる複数の文から 1 つの文を作成することは文融合と呼ばれている [1] が生じていることがわかる．表 7，8，9 にその例を示す．参照要約の中では 3 回この操作が行われており，入力文書における表現と比べ情報量を維持したまま文字数の削減が行われている．これらの操作によって削減された文字数を利用して参照要約作成者はさらに情報を要約に詰め込んでおり，この操作を行う機構を持たない要約器は再現率において劣後せざるを得ない．

表 7: 文融合の例 1

入力文書	本来なら改革２年目の今年が正念場となるはずである．ところが現実には，改革の熱気は薄い．
参照要約	本来なら改革２年目の今年が正念場となるはずだが，現実には改革の熱意は薄い．

表 8: 文融合の例 2

入力文書	例えば，朱首相が昨年公約した「８％成長の確保」は，７・８％に終わった．（中略）だが西側の経済専門家からは「本当は７・８％より低いのではないか」という疑問が出されている．
参照要約	公約の８％成長は７．８％だったが，本当はこの数字より低いのではないかという疑問が専門家からも出されている．

表 9: 文融合の例 3

入力文書	香港の繁栄回復が，中国の改革と切り離せないことを肝に銘じているのは中国のはずだ．にもかかわらず，中国の対応はあまりにも官僚主義的だった．
参照要約	香港の繁栄が中国改革と切り離せないことが，分かっているはずなのに，中国の対応は官僚主義的である．

ているため，これを利用できなかった．

5.3.3 文短縮・言い換え

表 3 を見ると，文書全体にわたって文の書き換えが行われていることがわかる．不要な修飾節などを除去する操作は文短縮と呼ばれており，この 表 3 の例でも文 1，文 10 などで典型的に行われている．一方，文短縮は典型的には係り受け木の枝狩りを通じて行われるが，参照要約に含まれる文のうち係り受け木の枝狩りによって実現できるものは少数であり，参照要約作成者はより洗練された操作を通じて参照要約を作成していることがわかる．

5.3.4 括弧の除去

表 3 の例において確実に行われている操作は括弧の除去で，括弧を通じて提供されている補足的な情報は全て要約から除去されていることがわかる．これによって大きく文字数を稼ぐことができるため，要約器もこの操作を実行できるようにする必要がある．

5.3.5 省略

便宜的に「省略」としたが,「この」や「など」の表現を用いて，入力文書における情報を除去している箇所がある．参照要約の文 3 では，朱首相の 3 つの実行のうち金融機構改革が失われており，これが「など」として表現されている．また文 6 では，改革と安定追求のジレンマを「この」で表現しており，同様に文字数を稼いでいる．

5.3.6 参照要約の信頼性

一方，参照要約の品質が疑われる部分もある．参照要約の 9 文め先頭の接続詞「また」は，入力文書を鑑みるに要約作成者の読みの誤りと思われる．

5.4 誤り分析の枠組みの適用

ここまでの分析を，本稿で提案した誤り分析の枠組みに適用した結果を表 10 に示す．表 10 に示されているように，今回は文短縮などの書き換え機構を利用していないため，非文が出力されることはなかった．一方で，文を短く書き換える操作を行えないため，情報の被覆において参照要約に大きく劣後しており，これが低い再現率の直接の原因となっている．

次に，分析の枠組みを自動要約の結果に適用する際の具体的な方法について表 11 に示す．表 11 では，あ

る誤りの種類がある誤りの原因によって生じる際に，どのようにそれが同定できるかをまとめたものである．自動要約の難しさの一因は誤りの調査の自動化が難しいという点にあるとも考えられ，その点を研究課題として取り上げ，ある入力文書，参照要約，自動要約の 3 つ組からこのような表を生成することを目標とした研究も考えることもできよう．

6 終わりに

本稿では，自動要約の誤り分析を扱った．一つの文書を取り上げ，それを要約器を利用して要約し，内部でどのような誤りが生じているか分析した．次に，筆者の考える自動要約の誤りの分類を提案し，それを利用して一つの文書の分析結果を分類した．また，どのような誤りが生じているかを調査するための具体的な方法についても提案した．特に今回取り上げた事例においては，以下の分析を通じて以下の結論を得られるものと思われ，それによって大きく要約の品質を向上させられるものと考えられる：

1. 話題の遷移を捉えそれを特徴量として要約器に認識させること．
2. 文融合，文短縮，括弧の除去などの操作を要約器が行えるようにすること．

今後は，提案した分類をより精緻化し，個別の分析事例を蓄積していく予定である．特に，「操作の不足」の内実，すなわちどのような操作が不足しているために自動要約が失敗しているか，および「情報の不足」，すなわち要約器に十分な情報がそもそも与えられているのかは今後より分析を行っていく必要がある．また，分析の結果を元に，要約器を改良し，改良の結果も合わせて報告する予定である．

参考文献

[1] Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, Vol. 31, No. 3, pp. 297–328, 2005.

[2] Harold P. Edmundson. New methods in automatic extracting. *Journal of ACM*, Vol. 16, No. 2, pp. 264–285, 1969.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop Text Summarization Branches Out*, pp. 74–81, 2004.

[4] Hans P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 22, No. 2, pp. 159–165, 1958.

[5] National Institute of Standards and Technology. The linguistic quality questions, 2007. `http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt`.

[6] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, Vol. 4, No. 2, pp. 1–23, 2007.

[7] Hitoshi Nishikawa, Kazuho Arita, Katsumi Tanaka, Tsutomu Hirao, Toshiro Makino, and Yoshihiro Matsuo. Learning to generate coherent summary with discriminative hidden semi-markov model. In *Proceedings of the 25th International Conference on Computational Linguistics (Coling)*, pp. 1648–1659, 2014.

[8] 佐久間まゆみ（編）. 文書構造と要約文の諸相. くろしお出版, 1989.

[9] 奥村学, 難波英嗣. テキスト自動要約. オーム社, 2005.

表 1: 自動要約の誤り分析

		非文章の出力	重要部同定の失敗	文意の歪曲
操作の不足		非文に関しては，文を生成する，あるいは書き換える機構が不十分であるため非文が生成される場合が該当するが，現在の書き換えは文短縮が主流であるため，非文の生成は操作の不足というよりは言語解析の失敗による．非文章については，機械要約に適切な談話構造を与える機構が不足している場合が該当する．	参照要約の作成者が行った操作を機械が再現することができず，そのため要約長の制約などから重要な情報を要約に含めることができなかった場合．例えば参照要約の作成者が略語化によって文字数を節約した場合，機械も同様の操作を行わない限り参照要約に到達できない．	ゼロ代名詞の復元を行う機構を要約器が備えていない場合が含まれる．その場合，機械が作成した要約に対して，入力文書とは異なる理解がなされる可能性が生じる．また，文章を構成する論理関係が入力文書と異なる読みを許すものになっており，読者が誤った結論に到達する場合が含まれる．
特徴量の不足	特徴量の設定不足	文の書き換え規則の不足に伴って必須格の格要素を誤って除去した場合など．談話構造に関する情報がなく，要約に適切な論理構造を与えることができない場合など．	入力文書の特定箇所が要約に含まれるべき重要な情報を含んでいることを，特徴量の設定の不足によって機械が理解できない場合．固有表現や評価表現などの情報が付与されていない場合など．	要約器の失敗というより言語解析器の失敗．
	言語解析の失敗	係り受け解析器が係り受け解析を誤った場合や，述語項構造解析器が述語項構造の認識に失敗した場合，談話構造解析器が談話構造の解析に失敗したなど．非文章が出力される恐れが高まる．	自然言語解析の失敗によって適切な特徴量を機械が取得できなかった場合．固有表現認識に失敗した場合など．	ゼロ代名詞を誤って復元した場合などが該当．入力文書の内容を要約が含意せず，致命的な誤りとなる．
パラメタの誤り		文の書き換え規則の適用順序が正しくなく，誤って必須格の格要素を削除してしまった場合などが該当．	ある特徴量が適切な重みを得ておらず，重要文として認定されるべき文がそう認定されなかった場合．訓練事例の不足や，不適切な学習方法が用いられた場合などが含まれる．	要約器そのものに起因するというよりは解析器の失敗，不備によるものが多いと思われる．
探索の誤り		パラメタは問題がないが，最適解が得られなかったために文の書き換えに失敗した場合．貪欲法などの近似解法を用い，最適解に到達できなかった場合が含まれる．	左に同じ．	左に同じ．
情報の不足		非文という意味ではこの場合は存在しないと思われる．	要約のために必要な情報がそもそも要約器に与えられていない場合．例えば，新聞記事のタイトルに含まれる情報を必ず要約に含めるように要約作成者が要約を作成しているにもかかわらず，要約器に対してはタイトルの情報が与えられない場合．	入力文書が本質的に曖昧性を含んでおり，外部の情報なしには入力を正しく解釈できない場合など．そのような場合，機械による解釈が誤り，結果として文意を歪曲した要約が作成される場合がありうる．

表 3: 文書番号 990305053 のテキスト．毎日新聞’99 データ集より引用した．太字は入力文書と参照要約とで文アライメントを取り，対応づけが取れた文同士において共通の単語である．

1. 中国の国会，全国人民代表大会（全人代）が ５日から始まる．
2. 朱鎔基首相の「政府活動報告」と予算案を審議し，私有制経済の存在を保障する憲法の一部改正などを行う予定だ．
3. 昨年 の全人代で， 新首相に選ばれた朱首相は ，「８％成長」と「三つの実行」（国有企業改革，金融体制改革，行政機構改革の３年以内解決）などを公約した．
4. 国有企業改革 ， 行政機構改革 は 計 二 千 数 百 万 人 規 模 の 大 リ ス ト ラ 計 画 で あ り，「命をかけてやる」と言い切った首相の 強い決意に称賛の声があがった．
5. 本来なら改革２年目の今年が正念場となるはず である．
6. ところ が現実には ， 改革の熱気は薄い．
7. アジア金融危機の影響が中国 に及び，経済環境 が急速に悪化し た．
8. 改革で生まれる失業者を他の産業に吸収できない．
9. 改革のテンポを緩めても，社会不安を抑え込むべき だという空気 が強まっている．
10. 安定追求とのジレンマがあっても 意志の強いことで知られる朱首相は改革路線を貫くと期待したい．
11. しかし， 中 国 の 経済 が 悪化 す る とともに， 中 国 に 対 す る 信 頼 を 揺 る が せ る よ う な 問題もいくつか発生している．
12. 例えば，朱首相が昨年公約 した「８％成長 の確保」は，７・８％ に終わった．
13. ほぼ８％であり，公約は達成されたとされた．
14. だ が 西側の経済 専門家から は「 本当は ７・８％ より低いのではないか 」 という疑問が出されている．
15. 電力消費量 や国内輸送量 が増え てい ないのに，国内総生産 （ＧＤＰ） が増えるのはおかしいと統計の公正さに疑問が出された．
16. 広東省など 地方の成長率が１０％を超えた の も水増しを疑われている．
17. 金融改革 については， 外資の 取り 扱いで大きく揺れて いる．
18. 昨年秋，突然倒産した広東国際信託投資公司（ＧＩＴＩＣ）の負債の処理について，「正規に登録された外資は返済を保証する」という中国政府の方針が，今年になって引っくり返った．
19. 外資は「貸手にも責任がある」と突き放された．
20. 各地方の国際信託投資公司（ＩＴＩＣ）にも同様の問題が飛び火している．
21. そこでも同じ方針が貫かれると，今後中国へ向かう勇気のある外資はなくなるかもしれない．
22. 香港に対 しても，最近の中国の 姿勢 は， 硬直した感じが否めない．
23. 香港人が中国国内でもうけた子供に香港居留権があると判断した香港の裁判所を，中国当局者が激しく批判した．
24. 「香港基本法」の解釈権は中国の全人代にある．
25. 一地方政府にすぎない香港の裁判所に解釈権はない，という趣旨だった．
26. 香港の最終審長官が，「全人代の解釈権を侵害する意図はない」と釈明して収拾された．
27. だが「１国２制度」に対する香港市民の自信はこの一件で急落した．
28. 中国の国有企業は，香港の株式市場で資金を調達する予定だった．
29. ところが香港の不況で，上場延期に追い込まれている．
30. 香港の繁栄 回復 が ， 中国 の 改革 と 切り離せないこと を肝に銘じているのは中国のはずだ．
31. にもかかわらず， 中国の対応は あまりにも 官僚主義的 だった．
32. 改革の直面する困難が大きければ大きいほど，柔軟な対応が必要になるだろう．

表 4: 文書番号 990305053 の参照要約

1. 中国の国会，全国人民代表大会が始まる．
2. 昨年，朱首相は「８％成長」と「三つの実行」などを公約した．
3. 国有企業改革や行政機構改革 など[省略] を「命をかけてやる」と言い切った首相の決意に称賛の声があがった．
4. 本来なら改革２年目の今年が正念場となるはずだが，現実には改革の熱意は薄い．[文融合]
5. アジア金融危機の影響で，中国でも経済環境が悪化し，社会不安を抑え込むべきという空気も強まっている．[文融合]
6. この[省略] ジレンマがあっても，朱首相は改革路線を貫くと期待したい．
7. しかし，経済悪化とともに，問題もいくつか発生している．
8. 公約の８％成長は７．８％だったが，本当はこの数字より低いのではないかという疑問が専門家からも出されている．
9. また[参照要約の信頼性] ，電力消費量が増えないのに，国内総生産が増えるのはおかしいと統計の公正さにも疑問が出された．
10. 地方の成長率が１０％を超えたことも水増しを疑われている．
11. 金融改革も，外貨の扱いで大きく揺れており，香港に対する姿勢も，硬直した感じが否めない．
12. 香港の繁栄が中国改革と切り離せないことが，分かっているはずなのに，中国の対応は官僚主義的である．[文融合]
13. 改革の困難が大きければ大きいほど，柔軟な対応が必要だろう．

表 5: 文書番号 990305053 の自動要約．太字は太字は入力文書と参照要約とで文アライメントを取り，対応づけが取れた文同士において共通の単語である．

1. **中国の国会，全国人民代表大会** （全人代）[括弧の除去] **が５日から始まる．**
2. 朱鎔基首相の「政府活動報告」と予算案を審議し，私有制経済の存在を保障する憲法の一部改正などを行う予定だ．[重要部同定の失敗]
3. **昨年** の全人代で，新首相に選ばれた **朱首相は** ， **「８％成長」と「三つの実行」** （国有企業改革，金融体制改革，行政機構改革の３年以内解決）[括弧の除去] **などを公約した．** [文短縮]
4. **国有企業改革，行政機構改革** は計二千数百万人規模の大リストラ計画であり， **「命をかけてやる」と言い切った首相の強い決意に称賛の声があがった．**
5. **本来なら改革２年目の今年が正念場となるはずである．**
6. **アジア金融危機の影響** が **中国** に及び， **経済環境が** 急速に **悪化し** た．
7. 改革のテンポを緩めても， **社会不安を抑え込むべきだという空気が強まっている** ． [文短縮]
8. 安定追求との **ジレンマがあっても** 意志の強いことで知られる **朱首相は改革路線を貫くと期待したい．** [文短縮]
9. **しかし，** 中国の **経済** が **悪化** する **とともに，** 中国に対する信頼を揺るがせるような **問題もいくつか発生している．** [文短縮]
10. 例えば，朱首相が昨年 **公約** した「 **８％成長** の確保」は， **７・８％** に終わった．
11. 香港人が中国国内でもうけた子供に香港居留権があると判断した香港の裁判所を，中国当局者が激しく批判した．[重要部同定の失敗]

表 10: 自動要約の誤り分析の一例

		非文章の出力	重要部同定の失敗	文意の歪曲
操作の不足		文短縮を利用しておらず非文の出力は行っていない.	文短縮や文融合といった，要約作成者がテキストに対して施した操作を模倣する機構を持たず，結果として低い再現率に甘んじている.	文意の歪曲を招くような文の組み合わせは生じていない.
特徴量の不足	特徴量の設定不足	上に同じ.	外資の取り扱いに関する話題の認定および香港に関する話題の先頭となっている文の認定に必要な特徴量を保持していなかったことは直接的に重要部同定に悪影響を与えている．また，文書末の文の重要度を低く評価したのは，文末の文をあらわす特徴量を設定しなかったことによると考えられる.	上に同じ.
	言語解析の失敗	上に同じ.	解析結果を見る限りは解析結果の誤りによる特徴量の抽出の失敗は生じていない.	上に同じ.
パラメタの誤り		上に同じ.	参照要約に含まれる特定の単語の重みを高めることによって要約器が正しい文を選択するように仕向けることができる．しかし，そのような重みの設定が広い事例に汎化されているかと考えると疑わしい.	上に同じ.
探索の誤り		上に同じ.	利用している探索ルーチンは最適解を保証しており，探索誤りは生じない.	上に同じ.
情報の不足		上に同じ.	入力した情報だけで要約器は参照要約に近づけるように観察される.	上に同じ.

表 11: 自動要約の誤り分析の枠組みの適用方法

		非文章の出力	重要部同定の失敗	文意の歪曲
操作の不足		入力文書と参照要約を比較し，参照要約の作成のために用いられている操作の同定．典型的には文短縮と文融合，括弧による補足情報の除去，略語化である．	一部の入力文を人手によって参照要約に含まれるものと同じものに書き換えることによって，当該操作を要約器が実行可能であった場合の重要部同定の性能を見積もることができる．	人手による入力文書および自動要約の読解．含意認識器の利用も考えることができるが，精度的に困難があろう．
特徴量の不足	特徴量の設定不足	文の書き換え規則などを追加し要約結果の変化を調査．	参照要約が含む重要な情報がどのような手がかりから得られるかを調査．固有表現，手がかり語など．	解析器の結果の確認．
	言語解析の失敗	解析器の結果の確認．	左に同じ．	左に同じ．
パラメタの誤り		人手によるパラメタの調整と結果の目視．パラメタを正しく設定することによって要約器が正しく動作する場合は，どのようにすればそのようなパラメタを推定できるか逆算する．	左に同じ．	左に同じ．
探索の誤り		整数線形計画問題ソルバーを利用し，最適解を得た上で，要約器の出力と比較する．	左に同じ．	左に同じ．
情報の不足		非文という意味ではこの場合は存在しないと思われる．	入力文書のタイトルや，挿入されている図表のキャプションなどが自動要約の際の重要な手掛かりとなっていないか確認する．それらをクエリとして与えクエリ依存要約とした場合の性能を調査する．	要約に与えられた情報のみに基づいて入力文書を適切に解釈できるか人手で確認する．外部の情報がない場合において，人間でも適切な読みが不可能である場合，機械でもそれは不可能である．

談話依存構造木に基づく要約手法の誤り分析

平尾 努
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

hirao.tsutomu@lab.ntt.co.jp

1 目的

本稿では，修辞構造木から得た談話依存木に基づく要約手法 [3] を対象として誤りを分析する．この手法で生成される要約は必ず談話依存木の根付き部分木となる．根付き部分木は EDU(Elementary Discourse Unit:ほぼ，節に相当) 間の依存関係をそこなわないため，テキストとしての一貫性が担保できる．しかし，根付き部分木という制約 (に加え要約長の制約) は非常に強い制約であるため，原文書から情報を抽出する際に大きな妨げとなることも考えられる．そこで，談話依存木が人間の生成する要約を再現するにあたり，どの程度の妨げとなるのかを議論する．

2 手法

2.1 要約手法

文献 [3] によると，談話依存木に基づく要約手法は以下の整数計画問題，木制約つきナップサック問題として定式化される．

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad \sum_{i=1}^{N} W(e_i)x_i \tag{1}$$

$$s.t. \quad \sum_{i=1}^{N} \ell_i x_i \leq L \tag{2}$$

$$\forall i : x_{\text{parent}(i)} \geq x_i \tag{3}$$

$$\forall i : x_i \in \{0,1\}, \tag{4}$$

$W(e_i)$ は i 番目の EDU の重要度，ℓ_i は i 番目の EDU の長さ (単語数) をあらわす．L は許容される要約長 (単語数)，x_i は i 番目の EDU を要約に含めるか否かあらわす決定変数である．関数 parent() は任意の EDU のインデックスに対し，談話依存木においてその親となる EDU のインデックスを返す関数である．式 (3) の制約があることで，要約が談話依存木の根付き部分木であることが保証できる．

表 1: それぞれの抜粋に対する評価結果

	抜粋 A	抜粋 B
ROUGE-1 w/	.501	.507
ROUGE-1 w/o	.451	.455
ROUGE-2 w/	.337	.342
ROUGE-2 w/o	.324	.321
再現率	.305	.316

2.2 コーパス

RST Discourse Treebank (RST-DTB) [1] に収録されている 30 件の要約 (人間が自由に生成した要約) に対し，その作成人物とは別の 2 名が独立に生成した抜粋を利用する．なお，抜粋は要約中の情報と原文書中の EDU を対応づけることで生成されている．

2.3 評価指標

自動評価手法の標準である ROUGE-N(N=1,2) のストップワードあり /なしに加え，参照要約が抜粋であることから EDU に基づく再現率を採用する．なお，システム要約に与える要約長の制約 L は参照要約 (抜粋) の単語数とした．

3 結果と議論

表 1 に ROUGE スコア，再現率を示す．なお，談話依存木は RST-DTB の人手による注釈付けからルール変換により得た．詳しくは文献 [3] を参照されたい．

参照要約は抜粋であるため，本来であればすべての評価指標は 1 を取り得るにもかかわらず，どのスコアとも決して高いとは言えない．この原因は，以下のいずれかであろう．

- 要約は根付き木に限るという制約 (式 (3)) に問題があった．

表 2: 木制約を考慮した場合のオラクルスコア

	抜粋 A	抜粋 B
ROUGE-1 w/	.825	.897
ROUGE-1 w/o	.819	.825
ROUGE-2 w/	.712	.723
ROUGE-2 w/o	.615	.617

表 3: 抜粋と談話依存木の関係

	抜粋 A	抜粋 B
根を含む割合	.633	.667
依存木正解率	.640	.662

- 根付き部分木の選択に問題があった (式 (1) の重み付けに問題があった).
- 上記，双方の問題があった.

より詳細に分析するため，木制約のもとで ROUGE のオラクルスコアを計算した．その結果を表 2 に示す．再現率のオラクルスコアは，簡単に計算することができないため，掲載していない．

表 2 より，根付き部分木 (式 (3)) という制約があることで，オラクルが 1 に満たないということがわかる．さらに表 1 のスコアは表 2 の 4〜6 割程度であることから，式 (1) の EDU の重み付けにも問題があることがわかる．つまり，要約は根付き部分木に限るという制約と EDU の重み付けの双方に問題がある．

EDU の重み付けに関しては，何らかの方策で改善できるだろうが，要約を根付き部分木に限るという制約が本当に問題なのだろうか? この疑問に答えるべく抜粋が談話依存木の根となる EDU を含む割合と抜粋がどの程度談話依存木を保存しているか示す指標である以下の依存木正解率を調べた．

$$\mathrm{DepAcc}(S) = \frac{\sum_{e_i \in S} \delta(e_i)}{|S|}. \qquad (5)$$

S は，要約として抽出された EDU の集合，$\delta(e_i)$ は，S に $e_{\mathrm{parent}}(i)$ が含まれる場合に 1，そうでない場合に 0 を返す関数である．

表 3 より，抜粋が根を含む割合，依存木正解率とも 0.6 程度であり，そもそも抜粋が談話依存木の根付き部分木の抽出によって生成されていないことがわかった．

これらの結果より，抜粋を再現するという観点からは要約を談話依存木の根付き部分木に限るという制約が悪影響を与えることがわかった．

4 結論

本稿では，文献 [3] の要約手法に対し，人手で生成した抜粋を再現するという観点から談話依存木の制約が及ぼす影響について分析した．その結果，要約を談話依存木の根付き部分木に限るという制約が悪影響を与えることが明らかとなった．しかし，この結果が要約手法としての限界を示しているとは限らない．

まず，制約を根付き部分木ではなく任意の部分木に緩和することで ROUGE，再現率が改善する可能性がある．文圧縮にて文献 [3] と類似性のある Katja らの手法 [2] では，単語依存木を用いて文圧縮をする際，依存木の根 (一般的には動詞) だけでなく，文中の他の動詞などいくつかの根の候補を考慮している．談話依存木は単語依存木のように文法的に規定された強い構造とは限らないため制約を緩める効果はより大きいのではないかと考える．

次に，本稿での評価の観点は抜粋を再現できるかという観点であり，要約 (抜粋) として人間が受容可能かという観点が欠けている．たとえ，ある人間が生成した抜粋を再現できなくとも，一定数の人間が受容可能な要約であれば工学的には問題なかろう．これら 2 点をより詳細に調べることは今後の課題としたい．

参考文献

[1] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pp. 1–10, 2001.

[2] Katja Filippova and Michael Strube. Dependency tree based sentence compression. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pp. 25–32, 2008.

[3] Tusomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *Proc. of the Empirical Methods in Natural Language Processing (EMNLP)-2013*, pp. 1515–1520, 2013.