記事に忠実ではない訓練事例も活用した見出し生成モデルの忠実性の改善法

植木滉一郎　平岡達也　岡崎直観
東京工業大学
{koichiro.ueki@nlp., tatsuya.hiraoka@nlp., okazaki@}c.titech.ac.jp

概要

見出し生成において忠実性の改善は重要な課題である．従来研究 [1] では，訓練データ中で忠実性が低い見出しを取り除くアプローチが提案されたが，見出し生成モデルの学習に用いる訓練事例数が減るため，生成される見出しの品質が低下するという問題があった．本稿では，マスク付き言語モデルを用いて忠実性の低い訓練事例を忠実性の高い事例に書き換える手法，タグにより見出しの忠実性を制御する手法を提案する．自動評価と人手評価の結果から，提案手法は既存手法と同程度の忠実性を保ちながら品質の高い見出しを生成できることを報告する．

1　はじめに

近年の言語生成モデルの発展により，見出し生成（記事に対して見出しを作文する）タスクの性能が向上し，人間が作成した見出しに近い品質の見出しを自動で生成できるようになった．しかし，自動見出し生成は記事内容から逸脱した見出しを生成することがある，という問題が報告されている [2, 3]．そのため，見出し生成モデルの忠実性の改善を目指した研究が進められている [4, 5, 6, 7]．

松丸ら [1] は，記事に忠実ではない見出しが訓練データの約 4 割に含まれていると指摘し，それが生成モデルの忠実性を低下させる要因になっていると報告した．忠実でない見出しを訓練データから取り除くことで，モデルの忠実性を向上させた．さらに，訓練事例の不足を補うため，自己学習を用いて擬似訓練データを作成した．

ところが，日本語データの実験において，自己学習で獲得した擬似訓練データはモデルの忠実性を改善するものの，参照見出し文との ROUGE スコアを低下させた．これは，訓練事例の見出しの中で，記事から逸脱している箇所は単語や句などの限られた範囲であるにもかかわらず，ゼロから擬似見出しを生成したことに起因する．そこで，記事から逸脱した見出しを破棄するのではなく，積極的に活用しながら見出しの忠実性と ROUGE スコアの両方を改善する手法を検討したい．

本稿では，マスク付き言語モデルを用いた訓練データの書き換え手法（図 1 中央）と，タグを用いてモデルの忠実性を制御する手法（図 1 右）を提案する．前者は，見出しの中で記事から逸脱している箇所の単語や句をマスク付き言語モデルで書き換え，擬似訓練データとする．後者は，訓練データには一切変更を加えず，モデルへの入力の先頭に忠実性の度合いを示すタグを付与し，見出し生成モデルを学習する．自動評価と人手評価の結果から，提案手法は既存手法と同程度の忠実性を維持しつつ，品質の高い見出しを生成できることが分かった．

2　提案手法

本研究の目標は，見出し生成モデルの ROUGE 値と忠実性を改善することである．具体的には，松丸ら [1] の手法をベースとし，より多くの訓練データを見出し生成モデルの学習に活用する手法を提案する．本稿では，**部分編集による擬似見出し生成**と，**タグによる制御**の二つの手法を提案する．

2.1　含意関係認識器

本研究では，記事に対する見出しの忠実性の判定を，記事と見出しの含意関係認識として扱う [1]．含意関係認識器を用いて，訓練データ $\mathcal{D}$ を忠実な見出し $\mathcal{D}^{\text{fai}}$ と，忠実ではない見出し $\mathcal{D}^{\text{hal}}$ のグループに分ける．含意関係認識器 E は，N 単語から構成される記事 $X = (x_1, x_2, \cdots, x_N)$ と，M 単語から構成される見出し $Y = (y_1, y_2, \cdots, y_M)$ の組を入力とし，記事 X が見出し Y を含意する確率 $E(X, Y)$ を予測す

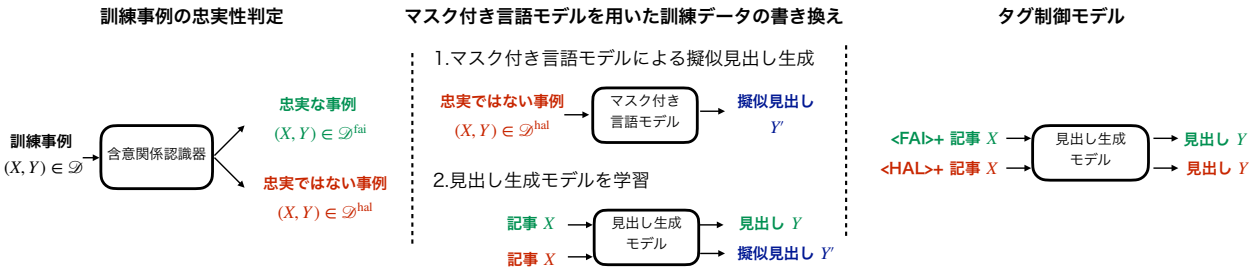

図 1　(左) 含意関係認識による訓練事例の忠実性判定 (中央) マスク付き言語モデルを用いた訓練データの書き換え (右) タグ制御による見出し生成

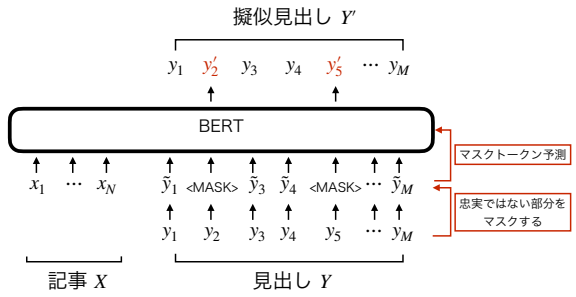

図 2　マスク付き言語モデルを用いた擬似見出し生成

る．$E(X,Y)$ が閾値 α 以上の事例を $\mathcal{D}^{\text{fai}}$ に，α 未満の事例を $\mathcal{D}^{\text{hal}}$ に入れる（図 1 左）．

2.2　部分修正による擬似見出し生成

マスク付き言語モデルで忠実ではない見出しを書き換え，忠実な擬似見出しを生成する（図 2）．まず，見出しの中で**記事に忠実ではない部分の検出**を行う．式 1 のように，記事と忠実ではない見出しの組 $(X,Y) \in \mathcal{D}^{\text{hal}}$ に対し，見出し Y に含まれるトークンのうち，記事 X に出現しないものを，忠実ではない部分と見なし，<MASK>トークンに置き換える．

$$\tilde{y}_i = \begin{cases} y_i & (\text{if } y_i \in X) \\ \texttt{<MASK>} & (\text{otherwise}) \end{cases} \tag{1}$$

続いて，**忠実ではなかった部分の穴埋め**を行う．式 2 のように，マスク付き言語モデル BERT [8] を用いて，マスクされた見出しから疑似見出しを生成する．具体的には，記事とマスクされた見出しを BERT に入力し，<MASK>トークンを予測することで，擬似見出し Y' を生成する．

$$Y' = \text{BERT}(X, [\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_M]) \tag{2}$$

このようにして，全ての $(X,Y) \in \mathcal{D}^{\text{hal}}$ に対して，見出しの書き換え $Y \longrightarrow Y'$ を行い，擬似訓練データ $(X,Y') \in \mathcal{D}^{\text{pse}}$ を構築する．最終的に，忠実な訓練データ $\mathcal{D}^{\text{fai}}$ と擬似訓練データ $\mathcal{D}^{\text{pse}}$ を用い，見出し生成モデルを学習する．

BERT モデルは，忠実な事例 $(X,Y) \in \mathcal{D}^{\text{fai}}$ の記事と見出しを連結して入力し，見出し部分のトークンの一部をマスクトークンに置換し，元のトークンを予測するタスクでファインチューニングする．

2.3　タグ制御モデル

訓練データの見出しには忠実性の高いスタイルと低いスタイルがあることを明示し，タグ制御により見出し生成の忠実性を制御するモデルを提案する（図 1 右）．生成を制御する既存手法 [9, 10] と同様に，訓練データの入力に忠実性を表す特殊トークンを連結する．具体的には，忠実な訓練事例 $(X,Y) \in \mathcal{D}^{\text{fai}}$ の記事 X の冒頭に特殊トークン<FAI>を，忠実ではない訓練事例 $(X,Y) \in \mathcal{D}^{\text{hal}}$ の記事 X の冒頭に特殊トークン<HAL>を追加する．

$\mathcal{D}^{\text{fai}}$ と $\mathcal{D}^{\text{hal}}$ は人手で作成された事例であるため，忠実度には差があるものの見出しの品質は高い．そこで，見出しの忠実度に応じた特殊トークンを訓練事例に付与し，見出し生成モデルを学習することで，訓練データの量を落とすことなく，忠実な見出しと忠実ではない見出しを書き分けられるようになることが期待される．このようにして学習した見出し生成モデルで忠実な見出しを生成するには，見出しを生成したい記事の冒頭に特殊トークン<FAI>を追加し，見出しを生成させればよい．

3　実験

3.1　実験設定

日本語の見出し生成タスクで実験する．訓練データとして，Japanese News Corpus (JNC)[1] を用いる．JNC は，朝日新聞の記事冒頭 3 文と紙面見出しの組から構成される約 180 万事例を収録している．また，評価用データセットとして，JApanese MUlti-Length Headline Corpus (JAMUL) を利用する．

1）https://cl.asahi.com/api_data/jnc-jamul.html

表 1　JAMUL データセットでの評価 (手法左のマークは図 3 に対応)

手法	訓練事例数	自動評価 R-1	R-2	R-L	含意率 (%)	人手評価 忠実 (%)	重要度	理解
● 全事例	1.7M	**49.15**	**21.20**	40.66	89.24	82	**4.13**	4.04
■ 忠実のみ [1]	0.8M	47.68	19.30	39.19	95.41	—	—	—
★ 忠実＋自己学習 [1]	1.7M	47.33	19.79	39.56	**95.80**	82	4.10	3.97
▲ 部分修正（2.2 節）	1.7M	48.46	20.76	40.44	95.14	80	3.98	3.97
◆ タグ制御（2.3 節）	1.7M	48.79	20.89	**40.75**	95.08	**84**	4.12	**4.07**

JAMUL は朝日新聞デジタルで配信された 1,524 件の記事全文と紙面見出し，10，13，26 文字以内の各種媒体向け見出しが付与されたデータセットである．本研究では紙面見出しを評価に用いる．トークン化には SentencePiece [11] を使用した．

日本語の事前学習済み BERT モデル[2)]を用いて，含意関係認識器を構築した．具体的には，松丸らが JNC のデータから作成した含意関係認識データセットから，5,033 件の訓練事例でファインチューニングを行った．1,678 件の評価データにおける正解率は 83.9%であった．この含意関係認識器を用いて，見出し生成の訓練事例を分類したところ，844,526 件 (全体の 48.9%) が忠実な事例と判定された．

提案手法の比較対象として，訓練データ全体で学習したモデル（全事例），松丸ら [1] の既存手法に従って忠実な事例のみで見出し生成モデルを学習したもの（忠実のみ），自己学習で擬似見出しを追加してモデルを学習したもの（忠実＋自己学習）を用いる．見出し生成モデルとして Transformer[12] を採用し，fairseq[3)]で実装した．

自動評価手法として，full-length F1 ROUGE 値を用いる．また，ROUGE では生成された見出しの忠実性を評価できないため [13, 14]，モデルが生成した見出しに対して，含意関係認識器（2.1 節）が含意と予測する割合（含意率）[1] を報告する．

人手評価では，忠実性（記事が見出しを含意するか），重要度（見出しは記事の重要な内容を含んでいるか），理解しやすさ（見出しは理解しやすいか）という 3 つの観点を用いた．「忠実性」は「はい」「いいえ」「判定不能」の 3 段階評価を行い，「はい」の割合いを報告する．「重要度」，「理解しやすさ」は，“5” を最も良いとする 5 段階評価を行い，その平均値を報告する．JAMUL のデータからランダムにサンプルした 100 事例に対し，各モデルで見出しを生成し，一人の被験者に評価を依頼した．

表 2　生成した擬似見出しの評価

生成手法	含意率 (%)	ROUGE-1	冗長さ (%)
元の見出し	0.00	100.00	8.1
忠実＋自己学習	91.01	43.45	4.6
部分修正	69.48	70.39	18.6

3.2　実験結果

表 1 に実験結果を示す．表中の R-1, R-2, R-L はそれぞれ，ROUGE-1, ROUGE-2, ROUGE-L を表す．両方の提案手法とも，忠実＋自己学習と同程度の含意率を達成しつつ，より高い ROUGE スコアが得られた．また，タグ制御モデルはシンプルな手法であるにもかかわらず，ROUGE スコアと含意率で安定した性能を示し，ROUGE-L 値では全事例を学習事例に用いた場合と同程度の性能を達成した．これは，提案手法が忠実性の問題に対処しつつ，より多くの訓練事例を有効に活用できたためと考えられる．

表 1 の人手評価によると，忠実性に関してはタグ制御モデル，重要度に関しては全事例を学習に用いるモデル，理解しやすさではタグ制御モデルが最も良い評価を得た．このように，タグ制御モデルは自動評価と人手評価の両方において良好な評価を得た．一方，部分修正モデルは人手評価による評価が芳しくなかった．部分修正モデルが生成する擬似見出しを確認してみると，冗長な繰り返し表現を含むことが多いことが判明した．これは，見出し中の複数のマスクトークンを予測するときに，重複した単語が予測されてしまうことがあり，結果的に重複を含む不自然な見出しがモデルの訓練データに用いられ，その性質を引き継いだ見出し生成モデルが学習されてしまったことが原因と考えられる．

2）https://github.com/yoheikikuta/bert-japanese
3）https://github.com/pytorch/fairseq

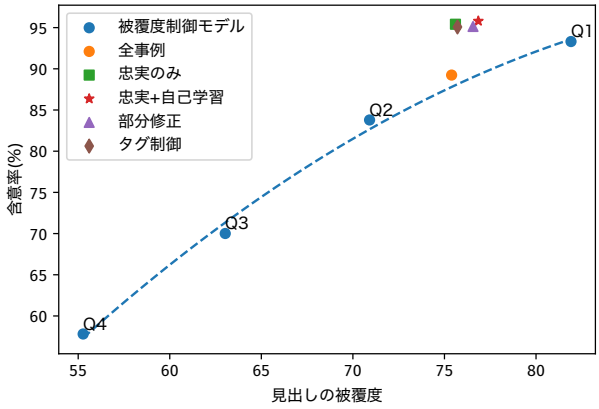

図 3　見出しの被覆度と忠実性に関する分析

4　分析

4.1　生成された擬似見出しの分析

マスク付き言語モデルによって生成された擬似見出しの忠実性と，書き換え前の見出しとの一致度，生成された見出しの冗長さを評価することで，擬似見出しの品質を分析する．同様に，既存手法 [1] である忠実＋自己学習によって生成された擬似見出しも分析する．擬似見出しの忠実性の評価には含意率を用いる．また，元の見出しと擬似見出しの一致度を測定するため，元の見出しと擬似見出しの間の ROUGE-1 F1 値を算出する．さらに，式 3 で擬似見出しの冗長さを求め，その平均値を報告する．

$$\text{冗長さ} := \frac{\text{見出し内に複数回出現する単語の数}}{\text{見出しの長さ（単語数）}} \quad (3)$$

表 2 より，忠実＋自己学習による擬似見出しが高い忠実性を示すことが分かる．部分修正による擬似見出し生成は，元の見出しの一部を書き換えるだけの手法であるが，忠実ではなかった見出しの 7 割を忠実なものへ書き換えることに成功した．

また，書き換え前の見出しと書き換え後の見出しの間で ROUGE-1 スコアを測定した結果から，部分修正により生成された疑似見出しは，元の見出しからの変更が少ないことが分かる．これは，忠実ではない訓練データのうち，利用できる箇所はできるだけ利用するという本研究の狙いが達成できていることを示している．冗長さに関しては，部分編集による擬似見出しが最も冗長であった．

4.2　提案手法による忠実性の向上

見出し生成では**被覆度**（見出しを構成する単語のうち，記事に含まれるものの割合）[15] の高い見出し，つまり記事からの抽出に近い見出しを生成するという手法でも，忠実な見出しを生成できると言われている．ところが，被覆度を単に高めるだけの手法は，言い換えや流暢さといった抽象型見出し生成の長所を打ち消してしまう．そのため，被覆度を高めることを狙わずに，忠実性を向上させることが望ましい．

Ladhak ら [16] は，忠実性を向上させる手法の有効性を評価する際に，被覆度の影響を取り除くべきと主張した．そこで，被覆度の影響を排除するため，モデル間で生成する見出しの被覆度を揃え，忠実性を比較する．生成する見出しの被覆度を制御できる見出し生成モデル（被覆度制御モデル）を学習し，被覆度と忠実性の関係を図 3 に曲線で描画した（モデルの詳細は付録 A に記載した）．この分析において，見出しの忠実性は含意率で測定している．

図 3 に各手法の見出しの被覆度と含意率をプロットすることで，手法の忠実性を有効に向上させたかを評価できる．図 3 の曲線は，見出しの被覆度の制御のみで達成できる忠実性を示しており，この曲線よりも上に位置する手法は，被覆度を高める以外の面で，忠実性を向上させていると評価できる．この分析によると，本研究で提案した両方の手法は，曲線よりも上側に位置し，全事例で学習したモデルと同程度の被覆度で，より高い忠実性を達成していることが分かる．従って，提案手法は見出し生成の忠実性を安易に向上させているのではなく，有効な手法であることが示された．

5　おわりに

本稿では，日本語の見出し生成において，忠実性と ROUGE スコアの両方を改善させる手法を提案した．本研究では，既存研究で破棄されていた忠実ではない見出しを活用することで，見出し生成モデルの性能向上を目指した．具体的には，見出しの部分的な修正による擬似見出し生成手法と，タグ制御による見出し生成の制御手法を提案した．

日本語の見出し生成タスクでの実験から，両方の提案手法は既存手法と同程度の含意率で，より参照見出しに近い見出しを生成できることが分かった．また，人手評価から，タグ制御による見出し生成モデルが忠実性，重要度，理解しやすさにおいてバランスの取れた手法であることが示された．さらに，提案手法は見出しの被覆度を単に高めるのではなく，見出しの忠実性を向上させることを確認した．

6 謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」により得られたものです.

参考文献

[1] Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. Improving truthfulness of headline generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1335–1346, Online, July 2020. Association for Computational Linguistics.

[2] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, 2018.

[3] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics.

[4] Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. **arXiv preprint arXiv:2104.04302**, 2021.

[5] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. Reducing quantity hallucinations in abstractive summarization. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 2237–2249, Online, November 2020. Association for Computational Linguistics.

[6] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. **arXiv preprint arXiv:2104.09061**, 2021.

[7] Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. Constrained abstractive summarization: Preserving factual consistency with constrained generation. **arXiv preprint arXiv:2010.12723**, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In **Proceedings of the Workshop on Stylistic Variation**, pp. 94–104, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[10] Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 864–870, Online, November 2020. Association for Computational Linguistics.

[11] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[13] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[15] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. **arXiv preprint arXiv:2108.13684**, 2021.

表 3 訓練事例の被覆度による分割

被覆度による分割	平均の被覆度
Q1	82.27%
Q2	68.20 %
Q3	57.57 %
Q4	40.09%

表 4 被覆度制御モデルの見出し生成の結果

制御タグ	R-1	R-2	R-L	含意率 (%)	被覆度
<Q1>	47.98	**20.36**	**39.96**	**93.31**	**81.91**
<Q2>	**48.20**	20.26	39.82	83.79	70.91
<Q3>	46.76	19.09	38.58	70.01	63.03
<Q4>	44.66	17.87	36.58	57.81	55.27

A 被覆度制御モデル

ここでは，4.2 節で使用した被覆度制御モデルについて説明する．

Ladhak ら [16] にならい，生成する見出しの被覆度を制御できるモデル (被覆度制御モデル) を学習する．この被覆度制御モデルを使用し，被覆度と忠実性の関係を曲線で描く．具体的には，2.3 節と同様に，制御タグによって異なる被覆度の見出しを生成し，そこから被覆度と忠実性の関係を推定する．

訓練事例を 4 つのグループに分割する (Q1 ~ Q4 とする)．表 3 は，分割したグループ内での被覆度の平均を示す．Q1 が最も高く，Q2,Q3,Q4 となるにつれて平均の被覆度が低くなる．

それぞれのグループに属する事例の記事の冒頭に，対応する制御タグ (<Q1>~<Q4>) を付与する．タグを付与したデータで見出し生成モデルを学習する．生成時には，生成したい見出しの被覆度を，入力の記事に付与するタグで制御できると期待する．

各制御タグを用いて見出しを生成した結果を表 3 に示す．<Q4>タグで生成した見出しは被覆度が最も低く，<Q1>タグで生成した見出しは最も高い．これより，タグによる被覆度の制御ができているとわかる．また，<Q1>タグで被覆度の高い見出しを生成することで，高い忠実性を達成できることもわかる．ROUGE 値に関しては，被覆度の制御だけでは，全事例を学習に用いるモデルやタグ制御モデルほどの高い値は達成できていない．

<Q1>~<Q4>の生成結果から，忠実性と被覆度の関係を推定したものが 4.2 節の図 3 となる．