NTCIR-14 QA Lab-PoliInfo Formal Run における Summarization Task の結果と評価

渋木 英潔 [†1]　木村 泰知 [†2†3]　乙武 北斗 [†4]　内田 ゆず [†5]　高丸 圭一 [†6]
阪本 浩太郎 [†1†7]　石下 円香 [†7]　三田村 照子 [†8]　森 辰則 [†1]　神門 典子 [†7†9]

[†1] 横浜国立大学　[†2] 小樽商科大学　[†3] 理化学研究所 AIP　[†4] 福岡大学
[†5] 北海学園大学　[†6] 宇都宮共和大学　[†7] 国立情報学研究所
[†8] カーネギーメロン大学　[†9] 総合研究大学院大学

1　はじめに

QA Lab タスクは，評価型ワークショップ NTCIR[1] において，現実世界における質問応答システムの実現を目指して開催されている．NTCIR-11 の QA Lab[1] を第 1 回として NTCIR-13 の QA Lab-3[2] まで，世界史の大学入試問題を対象として実施してきたが，4 回目の QA Lab-PoliInfo[2]では政治情報を対象とした質問応答に取り組んでいる．QA Lab-PoliInfo では，地方議会会議録コーパス [3, 4] を用いて，議員の発言に含まれる意見やその根拠や条件などを抽出し，関係性などを理解しやすいように整理して提示することを最終的な目標としている．

QA Lab-PoliInfo では、会議録中の発言が引用として与えられた場合にその引用箇所に該当する会議録の範囲を特定する Segmentation タスク、発言者の意図が誤解されないように要約する Summarization タスク、発言中の政治課題に対する意見と事実検証可能な根拠を分類する Classification タスクの 3 つのタスクを設定し、2018 年 11 月末に Formal Run を行った[3]。本稿では、Formal Run の Summarization タスクの結果を報告するとともに、その評価方法について考察する。

2　Summarization タスク

図 1 と図 2 に、入力として与えられる元文書と、その正解となる参照要約の例を示す。元文書は東京都議会会議録[4]を、参照要約は『都議会だより[5]』を用いている。図に示すように、正解はヘッドライン生成 [6] のように一文程度の長さであり、強く圧縮する必要がある。一般的なヘッドライン生成と異なるのは、図 2 に示すように、元文書に複数の議題が含まれており、それらをまとめて要約[6]する場合があることである。本タスクでは、図 1 のように議題が 1 つのものを single-topic、図 2 のように複数の議題を要約するものを multi-topic と定義した。問題数は、single-topic が 79 問、multi-topic が 67 問の計 146 問だった。

入力（元文書）

地産地消の 取り組みは 電力だけにとどまることなく、今こそエネルギーや水の自立的な都市東京を目指すべきです。他県の 犠牲、他県の 人々の 生活を 犠牲にしなければ、東京が 利水上も 治水上も 本当に 成り立たないのかどうか 検証が 必要と 考えます。今回の 福島原発事故後の 三月末には、放射性物質の 飛散により、水道水からも 乳児の 飲用基準を 超える 放射性沃素が 検出されました。その 後、非検出が 続いており、検査体制も 強化されているようではありますが、今後、台風や 豪雨等で 堆積していた 放射性物質が 川に 流され、水道水から 新たに 放射性物質が 検出されるおそれも 想定されます。水道水の 検査体制や 放射性物質の 除去等、水道水の 安全対策に どのように 取り組んでいるのか 伺います。また、地表に 堆積している 放射性物質、いわゆるホットスポットともいわれている 放射性物質汚染の 激しい 地域を 調査し、除染を 行うことが 今後の 重要な 課題でもあり、国の 決定を 待たずとも、区市町村と 連携して 取り組んでいただきたいと 要望いたします。震災以降、水道水に 関して 引き続き 検査が 継続され、結果も 報告されてはいますが、万が 一の ときのために も、災害時の 地下水の 活用について 注目すべきと 考えます。現在の 水道事業における 都内の 地下水の 利用実態について 伺います。さらには、新たに 利活用できる 地下水はないか、今のうちから 研究すべきと 考えます。災害時を 含めて 地下水を これまで 以上に 積極的に 活用すべきです。今回の 大震災を 教訓として、地下水の 活用を 推進すべきと 考えますが、見解を 伺います。地下水は、新潟中越沖地震で 断水が 続いた 後も 活用され、深刻な 水不足を 防いだ ともいわれています。厚生労働省が 行っている 原発事故後の 福島や 周辺の 地下水の 実態調査では、深井戸から 水を くみ 上げる 表流水の 影響を 受けない 地下水は、放射性物質による 汚染を 受けていません。しかし、現状では、地下水は 都の 保有水源として 位置づけられて おりません。利用実態が ありながら、課題を 抱える 水源に すら 位置づけられていない ことが 都の 利水計画における 現状です。この 現状を 改めて、地盤沈下や 地下水盆の 実態調査を 行った 上で、地下水を 災害時にも 有効な 貴重な 保有水源として 位置づけるべきと 要望いたします。

正解（参照要約）

災害時を含め積極的な活用推進を。

図 1: 元文書と参照要約の例（single-topic）

Formal Run 全体では 15 チームの参加があったが、Summarization タスクには、その内 7 チームが参加し、延べ 14 システム[7]からの提出があった。評価は、参加者による人手評価と ROUGE[7] による自動評価の 2 通りを行った。これは過去の QA Lab と同様の評価方法であるが、大学入試問題から地方議会会議録に対象が変わったことでどのような影響があるかは不明である。本稿では、この点の調査を行う。

3　人手評価

まず、146 問全てに対して、14 システムから提出された要約結果と正解（参照要約）を問題ごとにまとめたものを用意し、それを single-topic、multi-topic に関係なく 7 セット[8]に分割した。Summarization タスクに参加した 7 チームに対して 2 セットずつ割り当て、どの問題も 2 チームから評価されるようにした。各チームには以下のように評価を依頼した。

[1] http://research.nii.ac.jp/ntcir/index-ja.html
[2] https://poliinfo.github.io/
[3] Formal Run の各タスクおよびデータセットの詳細については、文献 [5] を参照されたい。
[4] https://www.gikai.metro.tokyo.jp/record/
[5] https://www.gikai.metro.tokyo.jp/newsletter/
[6] 参照要約では箇条書きとなっているが、タスクとしてはスタイルは指定していない。与えられた要約長の中で自由に記述してよい。
[7] QA Lab-PoliInfo では、任意のタスクに参加でき、1 つのチームが複数のシステムの結果を提出できる。
[8] 21 問が 6 セット、20 問が 1 セットとなった。

入力(元文書)

あわせて、ハッ場ダムが建設された場合に、治水の効果を発揮する地域はどこか伺います。ダムの洪水調節効果は、河道貯留効果といわれているように、下流に行くほど小さくなるとの指摘もあります。ダムに頼らずに、都内の河道整備、具体的には堤防整備、河床掘削、堤防強化等を行い、地域住民の安全を守ることができる現実的な治水対策を早急に進めるべきと考えます。ハッ場ダムが治水の効果を発揮すると想定される雨量と雨の降り方について伺います。自然の脅威に対して、ダムは本当に有効なのでしょうか。台風十二号の被害、ダムが人の命や財産を守るために機能しなかった事実が表面化したと、元国土交通省防災課長で淀川流域委員会委員長であった宮本氏は、記録的豪雨にダムは機能したかと題して、今回の災害直後に意見を述べております。ダムが効果を発揮するストライクゾーンは小さく、大災害にダムは極めて効果的には薄く、住民の命を守るためには、優先的にやらねばならないことがたくさんあるとも述べていました。ダムの効果に疑問を持ち、いかに住民の命を守るかといえば、第一には避難、避難体制を整えることであり、ダムをつくっても、その想定以上のものが来たら効果はなくなり、ダムができたから安心ではなく、自然というものはこれで終わるものではない、ダムをつくれば安全というダム安全神話から脱却しなければならないとのことでした。そこで、都は、集中豪雨や最近の台風などの豪雨にダムは機能すると考えているのか伺います。九月十三日、ハッ場ダム建設事業の関係地方公共団体から成る検討の場において、ハッ場ダム建設事業の総合的な評価案が発表されました。この発表について、ハッ場ダム建設容認の検証結果が出たとの報道がなされ、勘違いされている方もいらっしゃるようですが、当日の議題で、今後意見聴取の進め方も取り上げられているように、検証結果の中間発表にすぎず、最終的な評価はこれからということが正確な事実です。発表内容は、ダム事業推進にとって有利な結果となっており、これは、ダム事業を推進してきたダム事業者みずからが検証主体である上、推進を主張する関係自治体の意見のみが検証結果に反映され、ダムに疑問を持つ流域住民や有識者が意見を述べる場すら与えられていないからにほかありません。ハッ場ダムの検証作業について、事業の実施主体である国土交通省みずからが検証を行っていることの公平性と妥当性をどのように考えるか伺います。原発の安全管理を原発推進の経済産業省内の原子力安全・保安院が行ってきたこと、その結果、福島原発の事故が起きたことも考慮に入れるべきです。地震が頻発している福島の浜通り、双葉断層上に原発をつくることに警鐘を鳴らし続けていた専門家がいたにもかかわらず、保安院は昨年六月、福島第一原発双葉断層による地震の地震動評価を発表し、数地の地震動特性が十分に考慮され、不確かさについても適切に考慮されており、妥当なものと判断したという発表も行っていました。警鐘を鳴らす声は決して届かず、結果、今回の原発の大事故です。ハッ場ダムも同様、ダム建設が新たな災害を呼ぶ可能性について調査すべきです。ハッ場ダムの予定地では、地すべりの危険性等、専門家からの指摘もあります。こうした指摘について、都として ハッ場ダムの安全性をどう考えているのか伺います。適地ではない、安全性に問題がある場所にダムをつくってしまった例として、二〇〇二年に本体工事完成後、地すべりが起き、対策工事のため九年たったいまだに本格運用ができていない奈良県の大滝ダムがあります。地元住民の地すべりの懸念の声を無視し、万全の対策をとっているとして、ダム計画の見直しを行わなかったのは、事業者である国土交通省、当時は建設省です。国土交通省のいう万全の対策は万全ではないことが不幸にも証明されてしまっています。台風十二号の被害、大滝ダムの反省から ハッ場ダムは学ぶべきです。国土交通省の検証は中間結果であり、今後、ダム反対派の意見やパブリックコメントも実施した上で、最終的には、本体工事中止の判断、中止の英断が現政権で行われることを切に願い、質問を終わります。

正解(参照要約)

〔1〕治水の効果を発揮すると想定される雨量と雨の降り方は。〔2〕集中豪雨や最近の台風等の豪雨にもダムは機能すると考えているのか。〔3〕建設予定地は地すべりの危険性等の指摘もある。安全性に対する考えは。

図 2: 元文書と参照要約の例（multi-topic）

評価の観点は、「内容」、「表現」、「全体」の 3 通りで、「○」「△」「×」の 3 段階（「内容」のみ「▲」を加えた 4 段階）で評価した。

「内容」は、正解や元文書の内容が要約結果にどれだけ含まれているかの評価で、表現の適切さや文法的な正しさなどは考慮にいれないように指示した。評価の 4 段階は以下のように説明した。

内容（Content）

○＝正解の内容をほぼ網羅している
△＝正解の内容をそこそこ含んでいる
×＝正解の内容が含まれておらず、元文書の要約として不適切である
▲＝正解の内容とは異なるが、元文書の内容を要約している

補足　基本的に、正解の内容が含まれているかどうかで判断する。ただし、正解の内容ではないが元文書の要約としては適切な内容であるということも考えられるため、4 番目の評価として「▲」を用意した。

「表現」は、要約結果の表現や文法がどれだけ正しいかの評価で、内容に関しては考慮に入れないように指示した。評価の 3 段階は以下のように説明した。

表現（Well-Formed）

○＝文法的に正しい日本語である
△＝一部おかしな表現があるが理解できる
×＝日本語になっておらず理解できない

「全体」は、内容や表現を含めて、総合的に要約が元文書の要約としてどの程度適切かを評価するように指示した。評価の 3 段階は以下のように説明した。

全体（Total）

○＝元文書の要約として適切である
△＝元文書の要約としてまあまあである
×＝元文書の要約として不適切である

全ての問題に対して、「○」を 2 点、「△」を 1 点、「×」を 0 点として計算し、評価の観点ごとに平均化したスコアを求めた。「内容」における「▲」の評価は、正解として 2 点とした場合と、不正解として 0 点とした場合の 2 通りのスコアを計算した。

表 1 に人手評価のスコアを示す。all-topic の範囲が、single-topic と multi-topic を合わせた総合スコアである。なお、満点が 2 点であることに注意されたい。文内要約が要求されるタスクでありながら、平均して「表現」のスコアが 1.5 以上と良好な結果だった。一方で、「内容」や「全体」の平均スコアは 0.5 前後であり、今後の課題となった。また、「▲」を正解として評価したスコア（0.603）の方が不正解とした場合のスコア（0.423）よりも高い値となり、『都議会だより』以外にも正解となりうる要約が一定数存在することが分かった。このことがシステム間の評価にどのように影響するかの考察を 5 節で行う。

4　自動評価

本稿では、ROUGE スコアの計算に、GitHub[9] 上で公開されているモジュールを利用した。ただし、このモジュールは英語を想定しているため、アルファベットと数字以外をフィルターする機能を削除することで日本語に対応させた。形態素解析には MeCab[10] を用い、辞書には unidic-mecab 2.1.2[11] の短単位を用いた。

ROUGE には幾つかのバリエーションがあるが、どれが本タスクに適したものであるか自明ではない。そのため、本稿では、ROUGE-N1、-N2、-N3、-N4、-L、-W1.2、-SU4 の 7 通りの手法を用い、各手法に対し、再現率を用いた場合と F 値を用いた場合のスコアを計算した。また、計算する形態素列として、分割された原文の表現をそのまま用いた場合（以下、「表層形」）、原形に戻した形態素列を用いた場合（以下、「原形」）、内容語とみなした形態素のみを用いた場合（以下、「内容語」）の 3 通りを計算した。内容語とみなした形態素とは、(1) 品詞が「助詞」、「助動詞」、「感動詞」、「空白」、「補助記号」、「記号-一般」以外であること、(2) 動詞の場合は「為る」、「居る」、「成る」、「有る」以外であること、(3) 名詞の場合は、「所」、「為」、「くらい」、「の」、「事」、「物」、「積り」、「訳」以外であること、の 3 つの条件を満たした形態素とした。表 2 に all-topic を対象とした ROUGE スコアを示す。

[9] https://github.com/kylehg/summarizer/tree/master/rouge
[10] http://taku910.github.io/mecab/
[11] https://unidic.ninjal.ac.jp/

表 1: 人手評価のスコア（スコアの最大値は 2）

	all-topic				single-topic				multi-topic			
	内容		表現	全体	内容		表現	全体	内容		表現	全体
	▲=0	▲=2			▲=0	▲=2			▲=0	▲=2		
akbl-1	0.722	1.005	1.833	0.826	0.708	1.009	1.844	0.849	0.739	1.000	1.821	0.799
akbl-2*	0.707	1.000	1.837	0.793	—	—	—	—	0.707	1.000	1.837	0.793
KitAi-1	0.856	1.134	1.732	0.912	0.953	1.170	1.660	0.995	0.745	1.092	1.815	0.815
KitAi-2	0.788	1.035	1.308	0.667	0.849	1.028	1.340	0.722	0.717	1.043	1.272	0.603
KSU-1	0.043	0.043	1.955	0.048	0.052	0.052	1.934	0.057	0.033	0.033	1.978	0.038
KSU-2	0.076	0.121	1.745	0.071	0.080	0.156	1.722	0.104	0.071	0.082	1.772	0.033
KSU-3	0.091	0.157	1.715	0.104	0.104	0.179	1.731	0.156	0.076	0.130	1.696	0.043
KSU-4	0.111	0.167	1.419	0.093	0.118	0.193	1.420	0.132	0.103	0.136	1.418	0.049
KSU-5	0.048	0.078	1.692	0.048	0.057	0.085	1.726	0.057	0.038	0.071	1.652	0.038
KSU-6	0.078	0.169	1.535	0.091	0.085	0.151	1.542	0.094	0.071	0.190	1.527	0.087
LisLb-1	0.720	0.942	1.237	0.591	0.722	0.920	1.349	0.684	0.717	0.967	1.109	0.484
nagoy-1	0.886	1.104	1.619	0.899	0.953	1.179	1.642	1.028	0.810	1.016	1.592	0.750
TO-1**	0.504	0.846	1.763	0.551	0.464	0.794	1.778	0.521	0.550	0.905	1.746	0.586
TTECH-1	0.290	0.644	1.783	0.402	0.274	0.575	1.755	0.401	0.310	0.723	1.815	0.402
平均	0.423	0.603	1.655	0.435	0.387	0.535	1.532	0.414	0.406	0.599	1.646	0.394

*akbl-2 は single-type 未提出
**TO-1 はタスクオーガナイザによる結果

表 2: ROUGE のスコア（all-topic を対象）

		recall							F-measure						
		N1	N2	N3	N4	L	W1.2	SU4	N1	N2	N3	N4	L	W1.2	SU4
表層形	akbl-1	.400	.173	.113	.076	.345	.189	.157	.361	.156	.102	.068	.310	.167	.185
	akbl-2	.326	.124	.080	.057	.269	.147	.112	.320	.119	.077	.055	.262	.141	.144
	KitAi-1	.440	.185	.121	.085	.375	.217	.179	.357	.147	.096	.067	.299	.168	.188
	KitAi-2	.390	.174	.113	.078	.320	.200	.154	.343	.154	.101	.069	.281	.173	.176
	KSU-1	.158	.028	.009	.002	.147	.043	.071	.210	.039	.013	.004	.196	.059	.107
	KSU-2	.185	.043	.021	.014	.167	.063	.080	.230	.056	.027	.017	.209	.080	.116
	KSU-3	.172	.036	.008	.002	.157	.050	.075	.211	.043	.011	.003	.192	.062	.106
	KSU-4	.171	.044	.013	.002	.153	.055	.072	.219	.056	.017	.003	.195	.072	.106
	KSU-5	.227	.029	.010	.002	.195	.064	.089	.231	.029	.010	.003	.196	.065	.110
	KSU-6	.221	.038	.013	.004	.187	.065	.086	.230	.038	.012	.004	.192	.067	.108
	LisLb-1	.251	.120	.079	.058	.211	.132	.103	.226	.107	.071	.051	.188	.115	.118
	nagoy-1	.459	.200	.131	.089	.394	.229	.186	.361	.151	.097	.064	.305	.169	.192
	TO-1	.267	.093	.061	.045	.230	.117	.105	.272	.086	.052	.036	.233	.110	.133
	TTECH-1	.278	.060	.035	.020	.216	.092	.096	.240	.055	.031	.018	.187	.079	.111
原形	akbl-1	.415	.184	.122	.083	.357	.203	.164	.375	.165	.110	.074	.322	.179	.195
	akbl-2	.339	.135	.089	.064	.279	.158	.119	.333	.129	.085	.063	.272	.152	.153
	KitAi-1	.458	.199	.134	.096	.389	.234	.188	.373	.159	.106	.075	.311	.182	.199
	KitAi-2	.399	.179	.118	.082	.326	.208	.158	.351	.160	.106	.074	.286	.180	.181
	KSU-1	.161	.028	.010	.002	.148	.044	.071	.214	.040	.013	.004	.197	.061	.108
	KSU-2	.187	.044	.021	.014	.170	.064	.081	.233	.057	.027	.017	.212	.082	.117
	KSU-3	.175	.036	.008	.002	.159	.052	.075	.217	.044	.011	.003	.196	.065	.108
	KSU-4	.174	.045	.014	.002	.155	.056	.073	.222	.058	.018	.003	.197	.073	.107
	KSU-5	.230	.029	.010	.002	.199	.066	.090	.236	.030	.010	.003	.201	.067	.112
	KSU-6	.226	.040	.013	.004	.189	.066	.087	.235	.039	.012	.004	.195	.069	.109
	LisLb-1	.261	.125	.084	.061	.218	.139	.106	.235	.112	.075	.055	.195	.121	.122
	nagoy-1	.479	.217	.145	.101	.412	.247	.197	.377	.165	.108	.074	.319	.184	.205
	TO-1	.273	.097	.065	.048	.233	.121	.107	.277	.089	.056	.039	.236	.114	.136
	TTECH-1	.289	.064	.037	.022	.222	.097	.099	.251	.058	.033	.019	.193	.084	.114
内容語	akbl-1	.256	.113	.065	.034	.247	.124	.148	.224	.098	.056	.031	.216	.100	.158
	akbl-2	.200	.094	.051	.032	.189	.095	.109	.188	.089	.049	.031	.178	.087	.127
	KitAi-1	.285	.145	.090	.050	.278	.154	.180	.224	.115	.071	.042	.217	.107	.170
	KitAi-2	.254	.126	.083	.053	.247	.131	.156	.214	.109	.069	.046	.208	.106	.159
	KSU-1	.048	.001	.000	.000	.047	.007	.032	.059	.001	.000	.000	.058	.009	.043
	KSU-2	.069	.014	.000	.000	.067	.019	.043	.083	.015	.000	.000	.081	.022	.059
	KSU-3	.041	.002	.000	.000	.041	.007	.027	.050	.002	.000	.000	.050	.008	.036
	KSU-4	.050	.002	.000	.000	.048	.008	.031	.064	.003	.000	.000	.061	.011	.044
	KSU-5	.067	.002	.000	.000	.062	.013	.041	.063	.003	.000	.000	.057	.011	.043
	KSU-6	.053	.003	.000	.000	.051	.008	.034	.051	.003	.000	.000	.049	.009	.037
	LisLb-1	.171	.083	.044	.026	.160	.088	.106	.140	.068	.036	.023	.130	.065	.102
	nagoy-1	.326	.164	.094	.046	.315	.168	.201	.249	.123	.067	.036	.239	.110	.187
	TO-1	.116	.055	.035	.012	.111	.056	.070	.106	.042	.023	.011	.101	.042	.076
	TTECH-1	.088	.028	.015	.007	.082	.033	.050	.076	.024	.012	.006	.071	.027	.054

表 3: 「全体」と ROUGE スコアとの相関係数

	recall							F-measure						
	N1	N2	N3	N4	L	W1.2	SU4	N1	N2	N3	N4	L	W1.2	SU4
表層形	0.924	0.955	0.964	0.968	0.915	0.953	0.893	0.900	0.942	0.957	0.959	0.852	0.946	0.882
原形	0.928	0.959	0.968	0.972	0.918	0.956	0.900	0.912	0.950	0.965	0.968	0.866	0.954	0.894
内容語	0.943	0.957	0.948	0.920	0.939	0.952	0.926	0.942	*0.963*	*0.953*	*0.924*	0.937	*0.956*	*0.935*

5 考察

人手評価における all-topic の「全体」スコアが求めるべき評価であり、これを基本として ROUGE などのスコアがどのように影響しているかを考察する。我々は、本タスクにおける各システムの要約間の相対的な差が評価できればよいと考えており、影響を測る指標としてピアソンの積率相関係数を用いた。all-topic の「全体」との相関が強いスコアほど、評価に適切な指標であると考えられる。

最初に、要約すべき議題の数による問題の違いによって影響があるかを調査する。all-topic の「全体」と single-topic の「全体」との相関係数は 0.995、all-topic の「全体」と multi-topic の「全体」との相関係数は 0.991 となった。ただし、single-topic との相関では、akbl-2 が未提出のため、akbl-2 以外の 13 結果を用いて計算している。どちらも非常に強い正の相関を示したことから、要約すべき議題の数の違いが評価に与える影響は小さいと判断し、以降、all-topic の結果を基に議論する。

次に、「内容」と「表現」が「全体」に対してどのように影響を与えているかを調査する。「全体」と「表現」との相関係数は-0.046 となり、ほぼ無相関であった。「全体」と「内容」との相関係数は以下のように求めた。「『都議会だより』とは異なるが元文書の内容を要約している」にあたる「▲」の評点を正解と同じ 2 点として計算した場合の相関係数は 0.983、不正解の 0 点とした場合の相関係数は 0.979 となり、どちらも非常に強い正の相関を示した。このことから、「全体」の評価では「表現」ではなく「内容」が殆どを占めていると考えられる。3 節で述べたように『都議会だより』以外にも正解となりうる要約は存在するが、相関係数の比較（0.983 と 0.979）からその影響は僅かであり、『都議会だより』を正解として用いることに大きな問題はないと考えられる。

最後に、人手評価に最も近い ROUGE の手法が何であるかを調査する。表 3 に「全体」と各 ROUGE スコアとの相関係数を示す。全体的にどの ROUGE スコアとも非常に高い相関を示しているが、最も高い値となったのは、「原形」に戻した形態素列を用いた場合の再現率による ROUGE-N4（表中下線部）の 0.972 であった。表 2 に示すように、ROUGE-N の値が大きくなるほどスコアの絶対値は 0 に近づくため、過去の QA Lab では-N1（「内容語」による再現率）を基本的に参照していた。今回の結果においても、-N1 の中では「内容語」による再現率を用いた場合が最も高い相関（0.943）となった。

再現率と F 値を比較した場合、全体的に再現率を用いた方が高い相関を示した。しかしながら、「内容語」を用いた場合には、-N2、-N3、-N4、-W1.2、-SU4（表中斜字体）において F 値の方が高い相関を示した。したがって、不必要な「内容語」が含まれていないかを考慮しつつ、モダリティのような機能的な表現が再現されているかを考慮する評価手法を開発することで、より適切な評価が行える可能性がある。

6 まとめ

本稿では、NTCIR-14 QA Lab-PoliInfo の Formal Run における Summarization タスクの結果を報告し、その評価方法について考察した。人手評価では「内容」が「全体」の評価の殆どを占めており、『都議会だより』を参照要約として用いることに問題がないことが分かった。ROUGE による自動評価では、「原形」に戻した形態素列を用いた場合の再現率による ROUGE-N4 が人手評価と最も高い相関（0.972）を示した。また、ROUGE-N1 の中では、「内容語」による再現率を用いた場合が最も高い相関（0.943）を示した。

謝辞

本研究は JSPS 科研費 JP16H02912 および平成 30 年度国立情報学研究所公募型共同研究の助成を受けています。本タスクの設計等にご助言いただいた理化学研究所関根聡氏、東北大学乾健太郎教授に感謝いたします。会議録の著作権等にご助言いただいた情報セキュリティ大学院大学湯淺墾道教授に感謝いたします。

参考文献

[1] Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. Proceedings of the 11th NTCIR Conference, 2014.

[2] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-13 QA Lab-3 Task. Proceedings of the 13th NTCIR Conference, 2017.

[3] 筒井貴士, 我満拓弥, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則. 地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案. 自然言語処理, Vol. 21, No. 2, pp. 125–156, 2014.

[4] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ototake, Shigeru Masuyama. Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures, Coling 2016 workshop, The 12th Workshop on Asian Language Resources, pp.78–85, 2016.

[5] 木村泰知, 渋木英潔, 乙武北斗, 内田ゆず, 高丸圭一, 阪本浩太郎, 石下円香, 三田村照子, 神門 典子. NTCIR-14 QALab-PoliInfo の Formal run データセットの構築. 言語処理学会第 25 回年次大会発表論文集, 2019.

[6] 長谷川駿, 平尾努, 奥村学, 永田昌明. 文圧縮を活用したヘッドライン生成. 言語処理学会第 23 回年次大会発表論文集, pp. 8–11, 2017.

[7] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 workshop 8, 2004.