Kyutechコーパスにおける抜粋要約のアノテーションと分析

山村崇† 嶋田和孝‡
†九州工業大学大学院 情報工学府
‡九州工業大学大学院 情報工学研究院
{t_yamamura, shimada}@pluto.ai.kyutech.ac.jp

1 はじめに

対話要約は，複数人対話の理解や分析のための重要な研究課題の一つである．対話要約において，対話コーパスは要約生成の分析や構築に必要不可欠である．

本研究室では，意思決定タスクを対象とした対話要約コーパスとして，Kyutech コーパス [7] を構築し，作成したコーパスを Web ページ上で公開している[1]．Kyutech コーパスは，対話要約に自由に利用できる日本語のコーパスであり，現在までに 9 つの対話を収録し，各発話のトピックタグと正解要約 (参照要約) などのアノテーションを行っている．一方で，良質な対話要約コーパスを構築するためには，AMI コーパス [3] に代表されるように，対話データの拡充や，更なるアノテーションが必要であると考えられる．

本論文では，Kyutech コーパスを対象とした抜粋要約の正解データの作成と分析結果について述べる．重要な発話を取り出す抜粋型の要約は従来より行われており，このような正解データは，対話の結論となった発話群を分析するためにも重要である．また，構築した抜粋要約の正解データと参照要約を比較し，その分析結果について説明する．

2 Kyutech コーパス

本節では，Kyutech コーパスのタスク設定とアノテーションについて述べる．

Kyutech コーパスは，4 名の参加者による複数人対話を対象としている．4 名の参加者は，ある都市のショッピングモールの経営者であるという設定のもと，そのショッピングモールのレストラン街にあるレストランの閉店後に出店させるレストランを 3 つの候補の中から 1 つ選ぶというタスクについて議論する．議論の際には，出店候補店のレストラン情報，閉店するレストランの情報と閉店理由，レストラン街の既存店の情報，ショッピングモールのある都市の人口などの統計情報，さらに隣接する市町村の情報などが書かれた 10 ページほどの資料が準備されている．対話参加者はこの資料を 10 分間黙読した後に 20 分間の議論を行い，出店レストランを 1 つ決定する．現在 4 つのシナリオを用意しており 9 つの対話データを収録している．

書き起こしでは，日本語話し言葉コーパス (CSJ) の書き起こし基準 [9] を基にして，0.2 秒以上のポーズを転記単位の区切りとし，フィラーなどのタグ付けが行われている．また，CSJ コーパスの転記単位では 1 つの発話が複数に分割されるため，発話単位のアノテーションが行われている．9 つの対話データにおいて，転記単位では 4,509 発話 (発話単位のアノテーションでは 2,810 発話) が含まれている．

Kyutech コーパスでは，対話要約のためのいくつかのアノテーションを行っている．対話文中には様々な話題 (トピック) が頻出することから，トピックのまとまり毎に対話文を分割するテキストセグメンテーション [4] が対話要約の前処理として用いられている [1, 6]．テキストセグメンテーションによって対話中の発話群がトピック単位に分割されることで，対話文中のトピックをより適切に捉えた対話要約が可能になると考えられる．そのため，議論をトピック毎に分割するために，各転記単位に最大 3 つのトピックタグが付与されている．表 1 に，各転記単位に付与されるトピックタグを示す．

また，発話意図タグの付与については，既存の発話意図タグセット [2, 3] を用いて Kyutech コーパスへ適用し，より適切なタグセットの検討を行っている [11]．

参照要約作成は，AMI コーパス [3] のガイドライン (Abstractive Hand Summaries Guidelines (Scenario)) に準拠した[2]．このガイドラインに従って，句読点などを含め 250 字以上・500 文字以内で，「対話の

[1] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

[2] http://groups.inf.ed.ac.uk/ami/corpus/guidelines.shtml

タグ	説明
CandX	候補店 1 に関するトピック
CandY	候補店 2 に関するトピック
CandZ	候補店 3 に関するトピック
Cands	候補店 1〜3 の複数に関するトピック
Closed	閉店したレストランについてのトピック
Exist1	既存店 1 に関するトピック
Exist2	既存店 2 に関するトピック
Exist3	既存店 3 に関するトピック
Exist4	既存店 4 に関するトピック
Exist5	既存店 5 に関するトピック
Exist6	既存店 6 に関するトピック
Exists	既存店 1〜6 の複数に関するトピック
ClEx	既存店および閉店したレストランの両方に関するトピック
Mall	UBC モール全体についてのトピック
OtherMall	他の場所にあるショッピングモールに関するトピック
Location	モール内でのレストラン同士の位置関係に関するトピック
Area	地域や都市についてのトピック
People	ターゲットにする顧客についてのトピック
Price	価格に関するトピック
Menu	メニューに関するトピック
Atomos	雰囲気に関するトピック
Time	営業時間などに関するトピック
Seat	座席数や回転率などに関するトピック
Sell	売り上げに関するトピック
Access	交通手段などモールへのアクセスに関するトピック
Meeting	話を進めるための議事提案や最終決定部に関連するトピック
Chat	レストラン選択には直接関係ない雑談
Vague	前後の文脈からも何を言っているか分からない場合

表 1: トピックアノテーションに利用されるタグ

内容を知らない人が読んだ場合でも，議論の内容を理解できるような要約」という方針で対話毎に 3 つの参照要約が作成されている．

3　抜粋要約の構築と分析

本節では，Kyutech コーパスにおける抜粋要約を作成する手順とその分析について述べる．

3.1　抜粋要約のアノテーション

抜粋要約作成は，対話中の発話と参照要約のリンク付けを行うことで抜粋要約を作成した．具体的には，発話ごとに参照要約の要約文と関連があるかどうかの二値のアノテーションを行った．対話ごとに，本文中の発話がその対話の参照要約の要約文を含意する発話であれば，その発話と参照要約のリンクを作成して抜粋すべき重要文とした．

各対話の 1 つの参照要約に対して，リンク付けを行った結果を表 2 に示す．表 2 において，発話数はその対話に含まれる総発話数であり，リンク数は参照要約へのリンク付けされた発話数を示しており，割合は各対話のリンク数の割合を示している．リンク数の割合は，9 対話平均して約 23%であった．

また，参照要約を分析したところ，抜粋対象が対話中に存在しない要約文が全体の 19%存在した．図 1 に，

タスク ID	対話 ID	発話数	リンク数	割合 (%)
C1	0313	759	240	31.6
	0320	505	124	24.6
	0326	502	76	15.1
C2	0326	566	160	28.3
	0327	284	52	18.3
C3	0323	324	102	31.5
	0327	445	118	26.5
C4	0320	637	69	10.8
	0326	487	98	20.1

表 2: 各対話における参照要約への発話のリンク付け

・会議では、定食和屋の閉店後に出店させる店舗について議論された。
・ＵＢＣモールの不採算が原因で閉店することになったミスターＫの代わりの新規店舗を沖縄料理あたる、台湾ヌードル、ボノパスタから選ぶ会議が行われた。
・ＵＢＣモールのレストラン街において閉店した定食和屋に代わる新規店舗をラーメンかいぶつ、つけ麺ふうじん、ボノパスタの三件の中から選ぶ会議が行われた。
・今回閉店するミスターＫの代わりとなる新規店として沖縄料理あたる、台湾ヌードル、ボノパスタから選ぶ会議が行われた。
・その後、つけ麺ふうじんとボノパスタについてそれぞれの利点および欠点が議論された。

図 1: リンク付けされていない参照要約文の例

リンク付けされなかった参照要約文の例を示す．これらのほとんどが参照要約の冒頭の文であり，対話のタスクについて説明している文章であった．このような要約文は，対話中で議論された内容ではなく，対話のタスクや対話の状況について説明している文章であるため，対話中のどの発話にもリンク付けされなかったと考えられる．

3.2　参照要約と抜粋要約における分析

本節では，参照要約と抜粋要約の結びつきにどのような関係があるか分析するために，トピックタグを用いた分析について説明する．

参照要約には 3.1 節で構築した抜粋要約のリンク付けが行われており，各参照要約文のトピックはそれらの抜粋要約のトピックタグから構成されていると考えられる．そのため，抜粋要約のトピックタグの集合を正解トピックとし，参照要約文から洗い出したトピックタグと正解トピックタグを比較することで，リンク付けにおけるトピックの結びつきについて分析を行った．

参照要約の各要約文に対して，アノテータがその要約文に適切だと考えられるトピックタグの選定を行った．ある参照要約文に対して，アノテータが選定したトピックタグと正解トピックタグの例を図 2 に示す．図 2 の参照要約文において，アノテータが選定したトピックタグが，*Mall*，*People*，*Menu*，*Price* の 4 つのタグである．アノテータが選定した 4 つのトピッ

図 2: 選定したトピックタグと正解トピックタグの例

クタグはそれぞれ，参照要約文の「UBC モール」「客層」「メニュー」「値段」といった語と対応しており，表 1 の説明文と関連があるトピックタグが選定された．また，図 2 の参照要約文に対して，抜粋要約としてアノテーションされていた対話中の発話群がもつトピックタグは，*Mall*，*People*，*Menu*，*Price*，*CandX*，*Meeting*，*Sell* の 7 つのタグであった．アノテータが選定したトピックタグと正解トピックタグ (抜粋要約のトピックタグ) において，*Mall*，*People*，*Menu*，*Price* の 4 つのトピックタグが一致していた．Kyutech コーパスの 9 対話の参照要約に対してトピックタグの選定を行い，各対話における正解タグとの一致数を表 3 に示す．また，選定したトピックタグと正解トピックタグの一致精度を評価するために，表 3 より精度，再現率，F 値を計算した結果を表 4 に示す．

表 3 および表 4 より，アノテータが選定したトピックタグの半数以上が一致しており，正解トピックタグの半数以上を選定できていた．これより，参照要約と抜粋要約のリンク付けにおいてトピックタグは非常に重要な役割を持っていることが分かる．

今回，3.1 節で作成したアノテーションは各対話の 1 つの参照要約に行われたものであり，1 対話につき 1 時間半程度かかる作業であった．アノテーションでは，より多くのアノテータによる一致率を計算し，信頼性のあるアノテーションが必要であるが，複数のアノテータによる作業はコストの高い作業である．さらに，ある参照要約文がその発話に関係しているかをタグ付けするには，毎回すべての対話中の発話に目を通さねばならず労力のかかる作業である．一方で，今回選定したトピックタグは，要約文に「メニュー」といった語があれば *Menu* タグをつけるなど，要約文の表層で単純に判断できるものがほとんであった．そこで，参照要約文のトピックタグを選定して同一のトピックタグを含む発話文の提示するなどの手法により，より

タスク ID	対話 ID	選定タグ数	正解タグ数	一致数
C1	0313	22	28	17
	0320	31	29	22
	0326	16	14	12
C2	0326	22	28	19
	0327	19	13	11
C3	0323	22	22	15
	0327	26	24	14
C4	0320	19	15	11
	0326	23	18	13

表 3: 選定トピックタグと正解トピックタグの一致数

タスク ID	対話 ID	精度	再現率	F 値
C1	0313	0.773	0.607	0.680
	0320	0.710	0.759	0.733
	0326	0.579	0.733	0.647
C2	0326	0.682	0.682	0.682
	0327	0.750	0.857	0.800
C3	0323	0.864	0.679	0.760
	0327	0.565	0.722	0.634
C4	0320	0.579	0.846	0.688
	0326	0.538	0.583	0.560

表 4: 選定したトピックタグの一致精度評価

容易に抜粋要約のアノテーションを行うシステムなどが実現できると考えられる．今後は，このようなシステムを導入するなどして，より信頼性のある抜粋要約のアノテーションを行う予定である．

4 抜粋要約の予備実験

本研究の最終的な目標の一つは，複数人対話における生成型要約手法の構築である．そこで本節では，3.1 節で説明した抜粋要約のデータを用いて，抜粋要約の抽出実験について説明する．

対話理解の重要なタスクの一つとして，抜粋型要約 (重要文抽出) は従来より，議論中の重要な発話を推定するために研究されている．生成型要約手法の一つとして，対話文の重要な発話を抽出し，抽出した発話の統合や補完を行い要約を生成する手法が存在する [5]．また，我々はすでに別の雑談対話データを対象に抽出型の要約手法を提案している [8, 10]．そのため，3.1 節で説明した参照要約と紐づく抜粋要約を抽出することができれば，それらを基にして生成型要約が可能になる．本節では，その足掛かりとして作成したデータを利用とした重要文抽出について検証する．

本研究では，機械学習 (RandomForest, SVM) を用いて，抜粋要約の抽出実験を行った．機械学習による分類器では，Bag of Words の単語の頻出情報や，発

	RandomForest		SVM	
素性	Conv	Task	Conv	Task
Bag of Words	0.147	0.114	0.226	0.233
言語・非言語	0.256	0.271	0.249	0.229
全素性	0.119	0.140	0.298	0.274

表 5: 抜粋要約の抽出実験 (F 値)

話単体や発話間の関係性に関する特徴などの言語・非言語情報の素性 [8, 10] の計 17 の素性を用いて重要文の抽出を行った．

発話単体の特徴に関する素性

(話者，発話の長さ，高頻度単語の有無など)

発話間の関係性に関する素性

(発話者の連続性，直前の発話中の単語を有無など)

非言語情報に関する素性

(発話速度などの発話時間情報)

抽出した重要文と正解データである抜粋要約の精度，再現率から F 値を算出して抽出実験の評価を行った．また，leave-one-out 法を用いて 2 種類のデータセット (Conv:1 対話抜き，Task:1 タスク抜き) を構築し，得られた結果を平均して評価した．前者のデータセットは単純に 1 対話をテスト事例として残りの 8 対話を訓練事例とするのに対して，後者では残りの 8 対話の訓練事例の中でテスト事例と同じシナリオで行われた対話は除外して作成した．

実験結果を，表 5 に示す．表 5 に示すように，全体的に F 値が低い結果となった．また，2 種類のデータセットの作成方法についても大きな違いは見られなかった．このことから，より意味的な情報の素性の追加や対話データの拡充が今後の課題として挙げられる．

また，3.1 節で述べたように，参照要約文の中で抜粋要約のリンク付けができない要約文が約 19%もある．これらの要約文を生成するためには，対話中には明示的に表れない対話環境の状態を把握することが重要であると考えられる．

5 おわりに

本論文では，複数人対話要約コーパスを対象とした抜粋要約の構築手法について説明した．今後は，各対話の抜粋要約を精査し，順次 Web 上で公開する予定である．また，Kyutech コーパスでは動画や音声といった情報も収録しているため，これらについてもアノテーションを検討する．さらに対話データを収集し，コーパス自体の規模を大きくすることも今後の課題である．

謝辞

本研究の一部は科研費 26730176 の助成を受けたものです．

参考文献

[1] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Generating abstractive summaries from meeting transcripts. In *Proceedings of ACM Symposium on Document Engineering (DocEng '15)*, pp. 51–60, 2015.

[2] Harry Bunt, Jan Alexanderssona, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hosida, Volha Petukhova, Andrei Porescu-Belis, and David Trauim. Iso 24617-2 : A semantically-based standard for dialogue annotation, 2012.

[3] Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, Vol. 41, No. 2, pp. 181–190, 2007.

[4] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, pp. 562–569, 2003.

[5] Yashar Mehdad, Giuseppe Carenini, Frank W. Tompa, and Raymond T. Ng. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Natural Language Generation (ENLG - SIGGEN 2013)*, pp. 136–146, 2013.

[6] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of INLG 2014*, pp. 45–53, 2014.

[7] Kazutaka Shimada Takashi Yamamura and Shintaro Kawahara. The Kyutech corpus and the topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pp. 95–104, 2016.

[8] Yo Tokunaga and Kazutaka Shimada. Extractive summarization based on a combined method using several features for multi-party conversation. In *ACIS International Journal of Computer and Information Science*, Vol. 16, pp. 12–21, 2015.

[9] 国立国語研究所. 日本語話し言葉コーパスの構築法. No.124, 2006.

[10] 山村崇, 徳永陽, 嶋田和孝. 時間情報とテキストセグメンテーションに基づく複数人対話要約手法. 言語理解とコミュニケーション研究会 (NLC). 電子情報通信学会, 2015.

[11] 日野優登, 山村崇, 嶋田和孝. Kyutech コーパスにおける発話意図タグの設計と分析. 言語理解とコミュニケーション研究会 (NLC), 第 3 回自然言語処理シンポジウム. 電子情報通信学会, 2016.