The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

句構造解析とクラスタリングを用いた会話履歴の要約

Dialog Summarization using Phrase Structure Analysis and Text Clustering

星野　綾子　　細見　格
Ayako Hoshino　　Itaru Hosomi

NEC セキュリティ研究所
NEC Security Research Laboratories

Contact centers have massive amount of dialog records to be used to improve the quality of their service. This paper describes a method of summarizing a large number of dialogs on the same topic into a tree structure. The method consists of two steps: 1) summarization of each dialog with phrase structure rules, and 2) organizing dialogs into a tree structure using text clustering. A common oddness was observed among summarized dialogs, and this problem was mitigated by applying Multi-Sequence Alignment (MSA). With the proposed method, we were able to summarize real-life dialogs into a reasonably small tree with only two hours of rule writing labor. Also, applying MSA helped to reduce the number of nodes and led to a higher purity score.

1. 概要

コンタクトセンタには，日々大量に蓄積される会話履歴を活用し，オペレータの応対品質を上げるという課題がある．そのためには，多数の会話履歴を会話の流れが把握できる形で，一覧できるよう要約することが求められている．本稿では，大量の同一トピックに関する会話履歴を木構造のグラフに要約する手法を提案する．

手法は２段階から成り，１）各会話を構造解析の後，２）会話集合をクラスタリング，を行う．各会話の構造解析においては，文脈自由文法を用いて会話の構造を解析し，ラベルを付与する．その後に，同一層のノード群に対し，ラベル及び出現単語に基づきクラスタリングを行う．

提案手法により，会話の流れを保持した形で大量の会話履歴を要約することが出来る．また，本稿では，この手法により起こる会話要約上の課題を提示し，多重整列（Multi-Sequence Alignment）により問題を軽減する．

提案手法により，実データを２時間程度のルール作成の手間により，充分小さいサイズの木に要約することができた．また，多重整列により，全体ノード数の削減とクラスタリング結果の評価指標である純度の向上が見られた．

2. はじめに

コンタクトセンタ [*1] では，同一トピックに関する会話が多数行われ，記録されている．この会話履歴データを生かし，オペレータの応対品質を上げることは，コンタクトセンタ業者にとって重要な課題である．しかし，単純な文書要約では，セールストークや質問の出現有無や順番等，会話の流れを把握することができず，オペレータにとって具体的な改善策が得られない．大量の会話履歴を，会話の流れが分かるように集約する手法が必要である．

例えば，会話履歴中の連続する発話を，「本人確認」「注文受付」等，業務の単位にラベル付けし，複数会話の流れを木構造で把握できるようにすることは，会話履歴集合の要約として有用である．

連絡先: 星野　綾子，NEC セキュリティ研究所，〒 211-8666
神奈川県川崎市中原区下沼部 1753，TEL 044-455-8652，
a-hoshino@cj.jp.nec.com

*1　電話のみを扱うコールセンタに対し，メールやチャットによる顧客接点をコンタクトセンタと呼ぶ．

3. 先行研究

会話履歴を文書と見なすと，大量の会話履歴の要約には，複数文書分割技術の利用が考えられる．複数文書分割技術とは，文書内の要素同士の対応関係を考慮に入れ，複数の類似内容の文書を分割する技術である．Jeoug と Titov は，複数文書トピック分割技術を提案し，教師なしベイジアンモデルと単純な split-merge アルゴリズムを用いて，文書を全文書に共通のセグメントと各文書に独自のセグメントとに分割した [Jeong 10]．

また，コンタクトセンタを対象分野とした会話履歴を木構造に構造化する技術として，Bangalore と Stent の研究ある [Bangalore 09]．Bangalore と Stent の手法による木構造は，Dialog を根とし，Task，Topic/Subtask, Dialog Act, Pred-Args, Utterance, Clause を順に各層に配置している．

また，Luis と Cohen はオンラインフォーラム上の会話履歴から木構造を生成した [Louis 15]．彼らは PCFG と LCFRS の２つの異なる文法規則による構造解析を行い，スレッドをサブトピックに分割し，特に LCFRS により隣接しない同一サブトピックに関する投稿をリンクさせ，クラスタリングを含む３種の手法により，会話履歴を幅方向に時系列順にサブトピックが並ぶ木構造に整理した．

Bangalore と Stent，Luis と Cohen のいずれの手法も会話の流れを深さ方向に保持した木構造を作ることは出来ない．つまり，木構造を根から葉まで辿ることにより，ある会話を開始から終了まで再現する木構造を生成できない．そこで，本稿では，トピックのラベルを与えることができ，尚且つ，根から葉を辿ることにより会話を辿ることができる木構造を生成する手法を提案する．

4. 手法

提案手法のステップ１ (句構造解析による会話要約; PRS)，ステップ２ (テキストクラスタリングによる会話履歴要約; CLS) を説明する．また，提案手法により生じた課題に対し，多重配列 (MSA) を用いた緩和策を説明する．

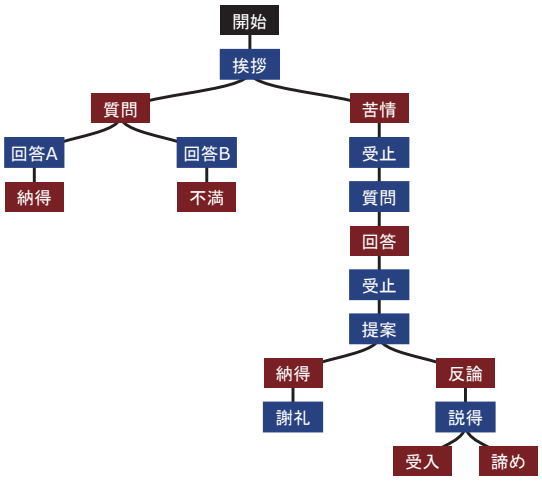

図 1: 会話構造化結果の例

4.1 句構造解析による会話要約

先行研究が示すように会話履歴データを纏め上げるのに句構造解析を用いることが適している．本稿では，解釈可能な要約結果を得るために，人手作成の句構造規則を利用し会話を句構造解析する方式を採用した．まず，文に対し,「質問」「回答」等の，発話行為に相当するラベルを付与する（非終端記号→終端記号規則）．実装には，正規表現パタンを用いた．次に，ラベル列をさらに要約するような規則を作成する（非終端記号→非終端記号規則）．これにより，複数文にまたがる事象を 1 ノードに要約する．実装には，独自のパタンマッチングプログラムを用いた．

規則作成の際には，コンタクトセンタにおける業務の目的（セールス，質問回答，等）を考慮し，目的に関連する事象をなるべく漏らさず捉えるようにする．以下に，作成した非終端記号→非終端記号規則の例を挙げる．O はオペレータの発話，C は顧客の発話を示す．

質問回答 → C:質問, O:回答
注文要求 → C:商品・数量指定
注文要求 → C:商品指定, C:数量指定

また，いずれの規則にもマッチしない文には「その他」ラベルを付与した．その他ラベルは，スペースの要求に応じて，連続するその他ラベルを 1 ノードに纏めたり，直前のその他以外のノードにマージすることもできる．

4.2 テキストクラスタリングによる会話履歴要約

ステップ 1 の結果，会話開始を根とし，深さ方向に会話の流れを保持する木を得た．さらに幅方向に複数ノードの纏め上げを行うことで，図 1 のようなラベル付の木を得る．

大量の会話履歴を要約するため，テキストクラスタリングを用い，同一親を持つ同一層における類似の内容のノードを統合する処理を，根から開始し再帰的に行った．尚，ノードは，そのノードに属する全ての文に含まれる一定頻度以上の形態素を保持している．アルゴリズムは，1) k-means または 2) 階層クラスタリング SLINK を採用し，実装には ELKI[*2] ツールキットを用いた．特徴には，形態素原形とラベルの BOW (Bag-Of-Words) を用い，統合されるノード群中の最頻のラベルを統合後のノードのラベルとした．ベースラインとして，同一親ノードを持つ同一ラベルのノードを纏める手法 (Simple) と比較した．

*2 https://elki-project.github.io/

表 1: 実データ処理時の木サイズ

	深さ	ノード数
before PRS	285.4	35,966
after PRS	20.9	2,407
after PRS-CLS	20.9	362

4.3 提案手法における課題と多重整列を用いた緩和

提案手法では，同一層のノード群を統合処理の対象としたため,「挨拶-注文要求-注文要求」等，同一のラベルが連続する木が多く見られた．これは，層 1 で 1 クラスタをなしたノード群のラベルとして「挨拶」が選ばれ，層 2，層 3 において 1 クラスタをなしたノード群のラベルとして「注文要求」が選ばれたためである．近接する層で同一のラベルが発生することが多いため，不自然に連続するラベル列が生じやすい．この課題を解決するために，ステップ 1 と 2 の間に，近接する層の同一ラベルノードが同一層となるようノード群を整列する多重整列の処理を加えた．多重配列を行うツールは，Clustal Omega[*3] を使用した．会話履歴群から得たラベル列をアミノ酸配列を表す記号に変換し，多重整列させた．整列の結果，挿入された空白に該当する個所に，空ノードを挿入した．

5. 評価

コンタクトセンタにおける実データについて，文法規則の作成時間，要約結果の木のサイズ，プログラム実行時間を評価した．また，多重整列を用いた要約の向上については，コーパスデータ [*4] により，木サイズと純度を計測した．

5.1 文法規則の作成時間

コンタクトセンタにおける実データ 126 件から，4 件程度のランダムサンプルを抽出し，2 時間程度で，非終端記号→終端記号規則 16 件，非終端記号→非終端記号規則 16 件，合計 32 件の規則を作成することができた．

5.2 要約結果の木サイズ，プログラム実行時間

ステップ 1，2 前後の木サイズを表 1 中に示す．要約前に比して，要約後はノード数がステップ 1 により 6.7%に削減され，ステップ 2 によりさらに 15%に削減された．実データで計測したプログラムの実行速度は，ステップ 1 は 2 分で処理した．また，ステップ 2 は 307 回クラスタリングを繰り返し，処理時間は 30 秒であった．要約結果は，末端が 45 本の木となり，紙やディスプレイ上で一覧可能な大きさとなった．

5.3 多重整列を用いた要約の向上

コーパスデータ処理時の木サイズ（ノード数）を表 2 中に示す．データセットに合わせて作成した，計 22 件の規則を用いた．木サイズを計測する際には，終端の空ノード及び空ノードのみを子孫として持つ空ノードは除外した．純度は，あるノードに統合されたノード群における，ノードに選出されたラベルと一致するラベルを持つノードの割合を木全体に渡って算出したものとする．

ステップ 2 に Simple を採用した場合, 多重整列によりクラスタリング後の木サイズは 117 になり，これは多重整列を行わない場合に比して 1.27%となった．尚，Simple を用いた場

*3 https://www.ebi.ac.uk/Tools/msa/clustalo/
*4 NTT 雑談対話コーパス [東中 14]

表 2: ノード数と純度

	Simple		Clustering	
	ノード数	純度	ノード数	純度
PRS	17,738	-	17,738	-
PRS-CLS	9,248	1.00	314	0.57
PRS-MSA-CLS	117	1.00	55	0.97

合には，純度は必ず 1.0 となる．また，多重整列を行うことにより木サイズが 17.5%に削減され，且つ，純度が 0.97 に向上した．

6. まとめ

本稿では，会話履歴を木構造に要約する手法として，句構造規則とテキストクラスタリングを用いた要約手法を提案した．実データに対し，句構造規則を作成することができ，コンタクトセンタの実データから，充分小さい木を生成できることを確認した．また，提案手法による課題の解決策として，多重整列を用いた改良を行い，ノード数の減少と純度の向上を達成した．

参考文献

[Bangalore 09] Bangalore, S. and Stent, A.: Incremental Parsing Models for Dialog Task Structure, in *EACL* (2009)

[Jeong 10] Jeong, M. and Titov, I.: Multi-document Topic Segmentation, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 1119–1128, New York, NY, USA (2010), ACM

[Louis 15] Louis, A. and Cohen, S. B.: Conversation Trees: A Grammar Model for Topic Structure in Forums., in Marquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y. eds., *EMNLP*, pp. 1543–1553, The Association for Computational Linguistics (2015)

[東中 14] 東中 竜一郎, 孝太郎 船越 : Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション, *SIG-SLUD*, Vol. B4, No. 02, pp. 45–50 (2014)