宿トピックの整理と自動分類の試み

林部 祐太　　István Varga
株式会社リクルート Megagon Labs, Tokyo, Japan
{hayashibe, istvan}@megagon.ai

概要

宿探し対話の自動応答に用いる知識整備として，ツリー構造での宿に関するトピックの整理に取り組んだ．トピック数（ノード数）は 791 個で，平均約 8 個のトピック例文をアノテーションした．また，発話の解釈結果をツリーにマッピングして自動応答に活用することを想定し，ツリーへの自動分類にも取り組んだ．自動分類には，トピック数が非常に多いが学習事例が非常に少ないという制約においても頑健に動くよう，類似度に基づく手法を用いた．

1　はじめに

宿を探しているカスタマーへの接客にて，オペレータは宿トピックに関する知識を用いて応対する．例えば次のような応対である．

- **要望が抽象的であれば，具体的な要件を例示して提案する（詳細化）**：
  カスタマー「自然豊かな所に止まりたい」
  →オペレータ「海が良いですか？山が良いですか？」
- **要望に該当する候補が多すぎれば，絞り込むために関連した追加の要望を聞く（要望追加）**：
  カスタマー「マグロを夕食に食べたい」
  →オペレータ「お食事はバイキングがお好みですか？それとも部屋食がよろしいでしょうか？」
- **要望に該当する候補がなければ，関連した代替となる要件を例示して提案する（代替提案）**：
  カスタマー「マグロとか食べられる宿はありますかね」
  →オペレータ「そのエリアでは，マグロをお出している宿はありませんが，サーモンの有名な宿があります．いかがでしょうか？」

これらは

- 「自然」に関するトピックには「海」や「山」などがある
- 食事に関係するトピックには「バイキング」や「部屋食」といった食事の形式に関するトピックがある
- 「マグロ」の兄弟トピックに「サーモン」がある

などといったような，宿トピックに関する知識に基づいて行われる．

我々はこのような応対ができる対話システムの構築を目指しており，本研究では宿トピックに関する知識の整理と，発話のトピック自動分類に取り組む．

2　関連研究

特定ドメインのトピックの整理を半自動で行った研究として，レストランレビューサイト Yelp に投稿されている和食，イタリアンなどといったレストランのカテゴリのレビューごとに，Latent Dirichlet Allocation (LDA)[1] を適用し，単語クラスタリングを行った研究 [2] がある．Reschke らは得られたクラスタをサブカテゴリとみなし，人手でクラスタトピックをアノテーションした．我々は経験による分類を反映させるため，自動クラスタリングは用いず完全に人手で整理する．

宿ドメインでは，次のようなトピックの分類で研究が行われている．Pontik らは SemEval-2016 Task 5 の宿のレビューの感情分析で 7 つのアスペクト (hotel, rooms, room amenities, facilities, service, location, food&drinks) を用いた [3]．Fukumoto らは 7 つのアスペクト（サービス，風呂，部屋，食事，立地，設備・アメニティ，総合）で宿のレビューを分類した [4]．安藤らは「商品の絶対的事実」「売り手の要望」「買い手の購入理由」など 23 項目で宿のレビューを分類した [5]．我々はこれらの分類を参考にしつつ，詳細化などの応対に使えるように，複数階層をもつ体系でトピックを整理する．

表 1　宿トピックとトピック例文の例

トピック	トピック例文
備品>アメニティ>アイマスク	アイマスクにこだわる
立地>交通アクセス>鉄道	鉄道でのアクセスが良い
風呂>水質>種類>死海風呂	死海風呂が最高だ
食事>アレルギー>そば	そばがだめだ
食事>専門>中華>餃子	餃子が出ました
食事>専門>洋食	洋食が食べられる

表 2　宿トピックツリー上でのトピックの分布

Level	1	2	3	4	5	6
節トピックの数	18	63	47	29	9	0
葉トピックの数	0	13	269	195	109	39

3　宿トピックツリーとトピック自動分類モデルの試作

3.1　宿トピックツリーの設計

先行研究の分類を参考にしつつ，旅行情報サイトじゃらん net[1] に掲載されている宿レビューを観察しながら，宿に関するトピックを手作業にてツリー構造で整理することにした．以下そのツリーを「宿トピックツリー」または単に「ツリー」とよぶ．そして，根の子を Level 1，その子を Level 2，. . . といったようによび，“Level 1> Level 2” といった表記でトピックを表す．

また，Level 1 と Level 2 はトピックの大分類・中分類を表すこととし，具体的なトピックは Level 3 以下で表すこととする．例えば「そばアレルギー」に関するトピックは，「食事>アレルギー>そば」という表記で表す．

Level 1 のトピックは，「立地，景色，宿タイプ，スタッフ，設備・サービス，インターネット，駐車場，風呂，食事，セキュリティ，備品，ベッド，部屋，ユニバーサルデザイン，値段，子ども，雰囲気，リゾート」の計 18 個とする．ツリー全体でトピックは計 791 個あり，節トピックは 166 個，葉トピックは 625 個である．トピックの例を表 1 に，分布を表 2 に示す．

3.2　学習用トピック例文の作成

自動分類器の学習のため，トピックに該当する発話文やレビュー文の例（以下，トピック例文とよぶ）を作成した．例を表 1 に示す．トピック例文は計 6,198 文で，平均約 8 文作成した．

1）https://www.jalan.net

3.3　トピック自動分類モデルの設計

宿トピック数は 791 と非常に多いが，学習事例数はトピックあたり平均約 8 個と非常に少ない．そのため，事前学習モデルに Softmax 層を追加して fine-tune する一般的な方法ではうまく動かない．そこで，類似度に基づく分類手法を用いる．

まず，トピック例文と入力を SentenceTransformers[2] [6] でベクトル化し，各トピック例文と入力のコサイン類似度を求める．次に，トピックごとにコサイン類似度の平均値を求め，それを入力に対するトピックのスコアとする．そして，閾値 t を超えるトピックのうち最も高いスコアをもつトピックを入力に対する予測トピックとする．ただし，閾値を超えるトピックが無ければどのトピックにも該当しないとする．

4　トピック自動分類実験

4.1　実験準備

事前学習モデルの準備

事前学習モデルにはじゃらん net に掲載されている宿レビューを使って学習した BERT[3] を用いる．BERT のトークナイザーは語彙サイズを 8,000 となるようして学習した SentencePiece[4] [7] を用いる．BERT のパラメータは公式サイトで公開されているモデル BERT-Base と同じように，バッチサイズは 512，Attention heads の数は 12，レイヤ数は 12，隠れレイヤ数は 12 とした．TPU を用いて 150 万ステップ学習し，Masked token prediction の精度は 66.5，Next sentence prediction の精度は 94.8 となった．

Sentence Transformers の学習とスコア計算の高速化

Sentence Transformers の学習は事前学習モデルとトピック例文を fine-tune することで行う．損失関数は同じトピックをもつ事例は近く，異なるトピックをもつ事例は遠くなるようなベクトルを得られる `BatchAllTripletLoss`[8] を用いる．学習は，バッチサイズ 128，エポック数 5 で行った．

処理の高速化のために，事前にトピック例文はベクトル化しておく．また，スコア計算は簡略化し，

2）https://www.sbert.net/
3）https://github.com/google-research/bert
4）https://github.com/google/sentencepiece

表 3　テスト事例の予測結果．予測トピックで括弧書きのものはスコアが閾値未満のため，該当トピックなしとする．

#	評価	入力文	正解トピック	予測トピック	スコア
#1	TP	海産物がメインの食事の宿を探してほしい。	宿タイプ>- 食事>食材>海鮮	食事>食材>海鮮	0.706
#2	TP	オーシャンビューの部屋が良い。	景色>から>部屋 景色>ターゲット>自然>海	景色>ターゲット>自然>海	0.870
#3	TP	駅からシャトルバスがある旅館で探してほしい。	宿タイプ>->旅館 立地>交通アクセス>駅 設備・サービス>サービス>送迎	設備・サービス>サービス>送迎	0.834
#4	TN	日付は 10 月 20 日から 4 泊だ。	-	(設備・サービス>サービス>アーリーチェックアウト)	(0.443)
#5	FP	食事代は予算をオーバーしても問題ない。	食事>値段>高め	食事>値段>安め	0.854
#6	FP	3 人ともお酒は 1 杯程度で十分な感じだ。	食事>内容>酒	食事>内容>酒>ビール	0.726
#7	FN	滞在して温泉にはいることが目的だ。	風呂>水質>温泉	(雰囲気>活動観点>リラックス)	(0.629)
#8	FN	エリアが千葉県だ。	立地	(立地>交通アクセス>高速道路)	(0.435)
#9	FN	早割りプランで探してほしい。	値段>付帯>割引	(値段>付帯>割引)	(0.558)

表 4　自動分類実験の精度

Level	TP	TN	FP	FN	Prec	Rec	F1
1	200	141	72	342	73.5	36.9	49.1
2	178	141	94	342	65.4	34.2	44.9
3	173	141	99	342	63.6	33.6	44.0
4	173	141	99	342	63.6	33.6	44.0
∞	173	141	99	342	63.6	33.6	44.0

表 5　自動分類実験の予測トピックごとの精度

トピック	TP	TN	FP	FN	Prec	Rec	F1	総数
風呂	20	5	3	2	87.0	90.9	88.9	40
備品	2	7	1	0	66.7	100.0	80.0	2
景色	6	2	1	2	85.7	75.0	80.0	10
ユニバーサルデザイン	7	4	0	4	100.0	63.6	77.8	13
設備・サービス	21	26	4	11	84.0	65.6	73.7	57
宿タイプ	14	19	1	12	93.3	53.8	68.3	190
ベッド	1	0	1	0	50.0	100.0	66.7	7
子ども	9	7	2	11	81.8	45.0	58.1	64
食事	40	19	19	46	67.8	46.5	55.2	128
部屋	20	8	14	23	58.8	46.5	51.9	67
駐車場	3	0	4	2	42.9	60.0	50.0	9
立地	25	6	10	46	71.4	35.2	47.2	155
雰囲気	2	8	1	6	66.7	25.0	36.4	20
値段	3	23	2	33	60.0	8.3	14.6	85
リゾート	0	4	2	6	0.0	0.0	nan	12
セキュリティ	0	0	0	0	nan	nan	nan	0
インターネット	0	2	0	0	nan	nan	nan	0
スタッフ	0	1	1	1	0.0	0.0	nan	3

高速近傍探索ソフトウェア NGT[5] を使って 50 個の近傍トピック例文とのコサイン類似度のみを利用して求める．

5）https://github.com/yahoojapan/NGT

4.2　開発データと評価データ

現在開発している対話システムでは，カスタマーの発話から省略されている言葉を補い，発話文を要望ごとに簡潔に解釈した解釈文を生成し，それを使って検索や返答を生成する．例えば，「それがいいですが，海鮮は出ますか？」といった発話文に対して「朝食はバイキングを希望する。朝食バイキングに海鮮が出るか知りたい。」といったような解釈文を作成する．そこで，解釈文に作成したトピックを自動分類して活用することを想定して実験する．宿探しの対話にそのような解釈文がアノテーションされたコーパス [9] のうち 1,443 文の解釈文[6] に対してトピックをアノテーションした．このうち 688 文を自動分類器のパラメータ開発に，755 文を評価に用いる．

開発用データにおいて，予測が正解であるとき[7] と不正解であるときのスコアの平均はそれぞれ 0.746 と 0.580 であった．そこで，閾値 t は平均の 0.663 とする．

4.3　自動分類実験結果

途中の Level まで一致していれば正解としたときの精度を表 4 に示した[8]．Level 1 のみの一致で評価した場合の F1 は 49.1 である．閾値未満はすべてト

6）論文中では「要約文」とよんでいる
7）1 つの文に複数のトピックがアノテーションされている場合は，1 つでも該当する場合「正解」とする
8）本論文では TP, TN, FP, FN はそれぞれ True Positive，True Negative，False Positive，False Negative の略とする

ピック無しとしているため FN が多くなり，Recall が 36.9 と低くなったと考えられる．Level 2 のみの一致で評価した場合の F1 は 44.9 である．Level 1 のそれと比較すると 4.2 ポイント低く，より詳細なトピックを予測するのは難しいことが分かる．予測トピックが完全に一致している場合のみ正解としたときの F1 は 44.0 であった．

表 5 には予測のトピックごとの精度を示した．「総数」はそのトピックが正解にある事例の総数を示す[9]．FN は「立地」，「値段」，「部屋」で比較的多いことが分かる．

予測の例を表 3 に示す．まず，#1 から#3 は TP の事例である．「海鮮」や「海の景色」の表現は学習事例に存在するが，「海産物」や「オーシャン」といった表現は学習事例には存在しない．しかしながら，ベクトルの類似度に基づくスコアにより適切に類似度を計算し，正しくトピックを予測できている．#4 は TN の事例で，どの学習事例とも類似しないことから，正しく該当トピックなしと予測できている．

#5 と#6 は FP の事例である．#5 は食事の値段という点では正しいが，高めの値段で良いという入力に対して逆のトピックを予測していため間違っている．#6 も食事の酒という点では正しいが，ビールに関する入力ではないので間違っている．

#7 から#9 は FN の事例である．#7 と#8 は最もスコアが高い予測トピックであっても間違っている事例である．#9 は閾値の設定によっては TP となった事例である．

5 考察・今後の課題

本論文での実験をもとに考察し，トピックツリーの改善すべき点と自動分類手法の改善すべき点について述べる．

5.1 トピックツリーの改善

今回の試作では，1 つのトピックの親は必ず 1 つとなるようなツリー構造でトピックを整理したが，複数の親が考えられるトピックは存在する．例えば，「沖縄そば」は「麺類」の子としたが，「郷土料理」の子とも考えられる．このようなトピックが存在するので，複数の親をもつことを許すなど，異なる構造を検討する必要がある．

また，トピックの網羅性を上げるため，実際の対話に適用し，実用上問題となるトピックが不足していないかを確認する必要もある．

5.2 自動分類の改善

自動分類の精度向上には，まず学習用データを改善することが必要である．学習に用いた例文は実際の発話をもとに作ってはいないため，学習と評価に用いたデータの種類が異なる．これを揃えることで精度の改善が見込める．ただし，すべてのトピックに対してアノテーションするのはコストが高いため，間違いやすいトピックに対してに集中して事例を追加するのが効果的だと考える．

次に，抜本的な精度改善のためにはモデルのアーキテクチャを改良することも必要である．現在のモデルでは「食事」トピックと「食事>値段」トピックのように親子関係があっても，`BatchAllTripletLoss` を用いているのでそれぞれに属すトピックは距離をとるように学習されている．これが#5 や#6 のような途中の Level までは正解しているが FP となっている事例の誤りの原因と考える．そこで，これを考慮に入れた損失関数へ置き換えることで改良できると考える．

また，ベクトルの分布の偏りが小さくなるような粒度の粗いトピックの分類には類似度に基づくスコアではうまく働かないのも問題であると考える．「立地」や「部屋」で FN が多かったのはそれらの粒度が他のトピックと比べて粗いためだと考えられる．そのため，類似度に基づくスコアとは別のスコアも用いることで FN を減らし，精度の改善ができると考える．

6 おわりに

本研究では宿探しの対話システム構築のために，宿トピックに関する知識の整理と，発話のトピック自動分類に取り組んだ．トピック数は非常に多い一方で学習事例は非常に少ない中，類似度に基づいた分類手法を用いた．

今後は作成したトピックツリーや自動分類器自体の改善はもちろん，対話システムに組み込んだ応用にも取り組んでいきたい．

謝辞 アノテーションを行っていただき，多くの示唆に富んだご意見をくださった山下華代氏に感謝します．また，有益な助言していただいた大阪大学の荒瀬由紀准教授に感謝します．

9) 1 つの文に複数のトピックが付与されうるので「総数」の和は文の総数とは一致しない．

参考文献

[1] David M Blei, Andrew Y Ng, et al. Latent Dirichlet Allocation. *The Journal of Machine Learning Research,*, Vol. 3, pp. 993–1022, 2003.

[2] Kevin Reschke, Adam Vogel, et al. Generating Recommendation Dialogs by Extracting Information from User Reviews. In *ACL*, pp. 499–504, 2013.

[3] Maria Pontiki, Dimitris Galanis, et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp. 19–30, 2016.

[4] Fumiyo Fukumoto, Hiroki Sugiyama, et al. Incorporating Guest Preferences into Collaborative Filtering for Hotel Recommendation. In *Proceedings of 6th International Joint Conference on Knowledge Discovery, Knowledge Enginnering and Knowledge Management*, pp. 22–30, 2014.

[5] 安藤まや, 関根聡. レビューには何が書かれていて、読み手は何を読んでいるのか? 言語処理学会年次大会, pp. 884–887, 2014.

[6] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *EMNLP*, pp. 3982–3992, 2019.

[7] Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *ACL*, pp. 66–75, 2018.

[8] Alexander Hermans, Lucas Beyer, et al. In defense of the triplet loss for person re-identification. *CoRR*, Vol. abs/1703.07737, , 2017.

[9] 林部祐太. 要約付き宿検索対話コーパス. 言語処理学会年次大会, pp. 340–344, 2021.