

# 対話行為スロット自動生成のための話題タグアノテーション

## Topic Tag Annotation for Dialogue Act Slot Induction

福永 隼也<sup>1\*</sup> 徳永 健伸<sup>1</sup> 西川 仁<sup>1</sup> 高橋 哲朗<sup>2</sup> 横野 光<sup>2</sup>  
Shunya Fukunaga<sup>1</sup> Takenobu Tokunaga<sup>1</sup> Hitoshi Nishikawa<sup>1</sup>  
Tetsuro Takahashi<sup>2</sup> Hikaru Yokono<sup>2</sup>

<sup>1</sup> 東京工業大学情報理工学院

<sup>1</sup> School of Computing, Tokyo Institute of Technology

<sup>2</sup> 株式会社富士通研究所

<sup>2</sup> Fujitsu Laboratories Ltd.

**Abstract:** We annotated a part of a real estate domain dialogue corpus with a set of topic tags corresponding to search condition attributes of a real estate database system. The annotation result of 50 dialogues revealed that we had a high inter-annotator agreement ( $\kappa = 0.78$ ) but many utterances could not be assigned the existing tags while containing useful information for search conditions. We also instructed the annotators to add ad hoc tags to those “out-of-tag” utterances. Our future plan is to develop a method to automatically derive additional tags related to search conditions regardless of dialogue domains.

## 1 はじめに

タスク指向対話システムの構築には、対話システムが発話内容を理解するために必要となる対話行為スロットの設計が重要である。レストラン検索などのデータベース検索タスクを行う対話システムの場合、データベース中の検索条件となる属性を対話行為スロットとしてそのまま用いる方法が考えられる。しかし、対話中には、直接の検索条件とはならないものの検索条件に関連する内容が出現する可能性があり、これらに対話システムで扱うためには対応する対話行為スロットを設計する必要がある。陽に与えられていない対話行為スロットを人手で設計することは人的・時間的コストが大きく、また作成したスロットを他のドメインにそのまま適用することは難しい。

そこで我々は、対話コーパスと検索対象のデータベースから、検索条件以外の対話行為スロットと、それらがどの検索条件と関係するかを、自動的かつ対話のドメインに依存しない方法によって出力することを目指している。そのための第一歩として、評価用データを作成すること、および検索条件に含められない内容が頻繁に出現することの確認を目的として、不動産ドメインの対話コーパスに対し、物件の検索条件を基とした話題タグをアノテーションした。

## 2 不動産対話コーパス

対話コーパスとして、横野ら [1] がクラウドソーシングによって作成した不動産ドメインの対話コーパスを用いる。このコーパスは、それぞれ不動産屋と客の役割を担う2名が、客の希望物件を探すためにおこなうチャット対話を収集したものである。不動産屋側は実際に物件を検索し提示することはせず、物件検索に必要な情報を得られたと判断した時点で対話を終了する。また、客側には10種類用意されたプロフィールのうち1つが割り当てられ、客側は割り当てられたプロフィールに合う条件の物件を希望する。各対話の1発話目は不動産屋側の「どのような物件をお探ですか」という内容の発話に固定されている。

コーパス中の対話数は986対話、総発話数は29,058発話、1対話あたりの平均発話数は29.5発話である。

## 3 話題タグ設計

不動産情報サイトSUUMO<sup>1</sup>の検索条件を基に、「賃料」、「間取りタイプ」など38種類のタグを作成した。これらはデータベース中の属性であると考えることができ、そのまま対話行為スロットとなりうる。また、検索条件に直接対応しない内容を発見することを目的として、これらのタグとは別に「その他」というタグを用意した。

\*連絡先：東京工業大学情報理工学院情報工学系  
〒152-8552 東京都目黒区大岡山 2-12-1  
E-mail: fukunaga.s.ab@m.titech.ac.jp

<sup>1</sup><http://suumo.jp>

表 1: 各アノテータのタグの使用数

アノテータ A		アノテータ B	
その他	230	その他	202
建物設備	129	建物設備	123
物件種別	119	賃料	108
賃料	106	物件種別	100
間取りタイプ	94	間取りタイプ	98

## 4 アノテーション

### 4.1 ガイドライン

アノテーションのガイドラインとして以下を設定した。

- 話題タグは発話毎に付ける
- 以降の発話は考慮せず、その発話とそれ以前の文脈のみから話題タグを判断する
- 該当する話題タグが存在しない場合は「その他」を付け、その内容を自由に記述する
- 1つの発話に複数のタグを付けることが可能であるが、以下のように場合に応じて記法を使い分ける（2つの記法を混在させることも可能である）
  - － 明示的に複数の話題を含む場合は、複数のタグを「+」によって結合する
  - － 話題の解釈が複数ありうる場合は、複数のタグを「/」によって結合する

例えば、「物件に何か必要な設備やセキュリティはありますか？」という発話は、物件の設備とセキュリティという2つの話題を明示的に含んでいるが、物件の設備には「室内設備」と「建物設備」の2つの解釈がある。この場合、この発話には「室内設備/建物設備+セキュリティ」というタグを付ける。

### 4.2 アノテーション結果の分析

アノテータ A と B の2名によって、不動産対話コーパス中の一部の対話に話題タグのアノテーションを行った。アノテーションを行った対話数は50対話、発話数は1,305発話である。客側のプロフィールごとの対話数は、10種類のプロフィールに対し各5対話である。

最初の発話は常に同じなので、これを除いた1,255発話について2名のアノテーション結果を分析する。

アノテーションの一致度をみると、話題タグが完全に一致したのは1,255発話中1,006発話（80.2%）であった。また、Cohen の kappa 値は0.78であり、2名のアノテーションは高い度合いで一致しているといえる。

表1に、各アノテータの話題タグの使用数を頻度の高い順に5つ示す。どちらのアノテータも「その他」の使用数が最も多く、検索条件に直接対応しない内容が無視できない頻度で出現することが確かめられる。

表2に、各アノテータが記述した「その他」の内容とその出現数を頻度の高い順に5つ示す。「その他」の内容は自由記述であるためアノテータ間での表記は異なり、その種類数はアノテータ A が29種類、アノテータ B が15種類であった。

表 2: 各アノテータの「その他」の内容と出現数

アノテータ A		アノテータ B	
住む人数	62	住む人数	56
希望	50	その他の希望	54
住環境	38	移動手段	13
住む人	8	物件の用途	12
交通手段	8	騒音	12

### 4.3 考察

今回作成したデータにおいて2名のアノテーションは高い度合いで一致しており、このデータを評価に用いることに問題はないと考えられる。

記述された「その他」の内容の多くは、直接は物件の検索条件とはならないが検索条件に関係するものであった。例えば、どちらのアノテータでも出現数が最多である「住む人数」は検索条件には存在しないが、住む人数が多ければ広い物件が必要となるという推論から「部屋の広さ」や「間取りタイプ」などの検索条件に関係するといえる。このような「その他」と検索条件との間の関係を捉えることが出来れば、直接検索条件を指定しなくても検索が行える、より柔軟な対話システムを作成できると考えられる。

また、「その他」の内容においては物件の属性やそれに関係するもの以外の内容が出現していた。例えば、ローンなどの「支払い方法」という内容が両アノテータによって記述されていたが、これは物件の属性には関係しない。対話のドメインが物件の購入や賃貸を前提とした検索であることから、購入というタスクに関する内容が出現したと考えられる。複合的なドメインについての内容も対話システムは扱える必要があるため、対話行為スロットとして抽出することが求められる。

## 5 おわりに

本研究では、検索条件以外の対話行為スロットを自動的に作成するための第一歩として、不動産対話コーパスに対し話題タグのアノテーションを行った。その結果、アノテーションは高い度合いで一致した。また、不動産対話には検索条件に直接対応しない内容が多く出現すること、および、それらの多くは検索条件に関係していることがわかった。

今回はコーパス中の986対話のうち50対話にのみアノテーションを行ったが、現在、残り936対話のアノテーションを進めている。残りの対話のアノテーションの完了後、同様の分析を行う予定である。また、今回作成したデータをテストデータとして、まずは各発話ごとの検索条件に対応するかの分類タスクを行うことを予定している。

## 参考文献

- [1] 横野光, 高橋哲朗. クラウドソーシングを用いた対話コーパス構築. 言語処理学会第23回年次大会発表論文集, pp.783-786, 2017.