Vol.2018-NL-236 No.12
2018/7/10

タスク指向対話におけるユーザ要求の理解とその根拠の抽出

福永 隼也[1,a)]　西川 仁[1,b)]　徳永 健伸[1,c)]　横野 光[2,d)]　高橋 哲朗[2,e)]

概要：本論文は，データベース検索を行うタスク指向型対話を対象として，ユーザ発話中で明示的に述べられていないユーザ要求の解釈をおこなう手法を提案する．ユーザ発話において，検索条件としてデータベースフィールドとその値が明示的に指定されない場合，その発話を直接データベースへのクエリに変換することはできない．しかし，そのような発話中にも明示的に述べられないユーザ要求が含まれる場合があり，それを解釈することは，対話システムがより自然で効率的なデータベース検索対話をおこなうために重要である．本論文ではこのように明示的に述べられないユーザ要求を非明示的条件と呼ぶ．また，非明示的条件の解釈を，ユーザ発話を関連するデータベースフィールドに紐づけ，また同時にその根拠となるユーザ発話中の文字列を抽出する課題として定式化する．このような新しい課題を提案するとともに，課題に対する 2 つの手法として，サポートベクターマシンに基づく手法と，分類と根拠となる部分文字列の抽出を同時に行うニューラルネットワークによる手法を実装した．不動産に関する対話のコーパスを利用した評価実験の結果，サポートベクターマシンに基づく手法がより良好な結果を示すことがわかった．

Interpretation of Implicit Conditions in Database Search Dialogue

FUKUNAGA SHUN-YA[1,a)]　NISHIKAWA HITOSHI[1,b)]　TOKUNAGA TAKENOBU[1,c)]　YOKONO HIKARU[2,d)]
TAKAHASHI TETSURO[2,e)]

1. はじめに

対話システムがユーザ発話から抽出するべき情報は，背後にあるアプリケーションによって異なる．対話システムをデータベース検索のための自然言語インタフェースとして用いる場合，対話システムはデータベースへのクエリを作成するために，ユーザ発話中で検索条件として指定されるデータベースフィールドとその値を抽出する必要がある．データベース検索対話において，ユーザ発話中からこのような情報を抽出する研究はこれまで多くなされてきた．例えば，文献 [1], [2], [3] は，ATIS (The Air Travel Information System) コーパス [4], [5] を用いて，ユーザ発話からデータベースフィールドの値を抽出する研究を行っている．ATIS コーパスは Wizard-of-Oz によって収集された，ユーザと航空交通情報システムとの対話コーパスであり，各ユーザ発話中の表現には，出発地や到着日などのデータベースフィールドに対応するタグが付与されている．しかし，実際の対話には，データベースフィールドには直接対応しないものの，クエリを作成するために有用な情報を含む発話が出現する．対話システムがこのような情報を利用することで，より自然で効率的なデータベース検索を行うことが可能である．例えば，不動産の検索対話において，家族の人数は物件の広さを決める上で有用な情報である．しかし，家族の人数は物件の属性ではなく客の属性であるため，不動産データベースには含まれない．我々はこのように，データベースフィールドには直接対応しないものの，データベース検索をおこなう上で有用な情報を**非明示的条件**と呼ぶ [6].

非明示的条件を利用する対話システムを実現するためには，以下の 2 つの課題に取り組む必要がある．

(1) 非明示的条件を含むユーザ発話から，データベース

1 東京工業大学 情報理工学院
Tokyo Institute of Technology, Meguro, Tokyo, Japan
2 富士通研究所
Fujitsu Laboratories Ltd., Kawasaki, Kanagawa, Japan
a) fukunaga.s.ab@m.titech.ac.jp
b) hitoshi@c.titech.ac.jp
c) take@c.titech.ac.jp
d) yokono.hikaru@jp.fujitsu.com
e) takahashi.tet@jp.fujitsu.com

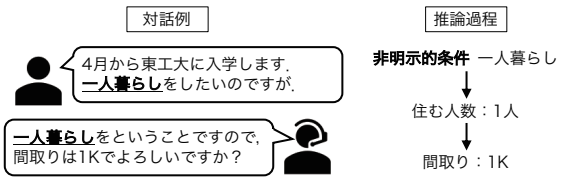

図 1 対話と推論過程の例

フィールドとその値の組を抽出する．

(2) ユーザ発話中から，(1) でおこなったデータベースフィールドと値の抽出の根拠となる部分を抽出する．

1 つ目の課題は，非明示的条件を含む発話からデータベースへのクエリを作成するために必要な処理である．2 つ目の課題によって抽出された根拠は，システムがユーザへの確認発話を生成する際に役立つ．データベースフィールドとその値の組を，非明示的条件から抽出するためには推論が必要となるため，その結果が常に正しいとは限らない．したがって，システムの解釈が正しいかどうかをユーザに確認することが望ましい．この際，システムがおこなった解釈の根拠を提示することで，より自然な確認発話を生成することができる．例えば，図 1 のやり取りにおいて，「一人暮らしをしたいのですが.」というユーザ発話から，システムが “間取り” というデータベースフィールドと “1K” という値を抽出したとする．このとき，単に「間取りは 1K でよろしいですか？」と確認するよりも，「一人暮らしということですので，間取りは 1K でよろしいですか？」とシステムが判断した理由を追加することでより自然な対話となる．このような場合，ユーザ発話中の「一人暮らし」という表現を “間取り” というデータベースフィールドとその値の根拠として抽出することで，確認発話の生成に役立つ．

また，「一人暮らし」を根拠として抽出することで，「一人暮らし」という非明示的な条件が「1K」という明示的な条件に結びついているという推論の過程を明らかにすることができる．そのため，例えば，顧客の住人構成などの情報が頻繁に対話に出現するようであれば，その傾向に基づいてシステムの運営者は当該項目をデータベースフィールドに新規に追加するなどの改良を施すことができる．また，推論の過程の可視化は，システムのアカウンタビリティのためにも重要である．

我々は，1 つ目の課題に対しては，まずデータベースフィールドの抽出をおこない，その後，値を抽出するという 2 つの段階に分けて取り組む．本論文では前段の処理を，ユーザ発話をデータベースフィールドへ分類する問題として定式化する．1 つのユーザ発話から複数のデータベースフィールドが抽出されることもあるため，この分類問題は，マルチラベル分類問題として定式化される．また，文献 [7] で提案された手法を拡張することで，データベースへの分類と，その分類の根拠の抽出を同時におこなう手法を提案する．後段の処理である値の抽出については，データベースの構造や内容を参照することが必要となるため，今後の課題とする．

2. 関連研究

伝統的なタスク指向対話システムは Natural Language Understanding (NLU), Dialogue State Tracking, Policy Learning, Natural Language Generation の 4 つのモジュールのパイプラインによって構成される [8]．NLU はさらに，ユーザ意図推定とスロットフィリングの 2 つの処理にわけることができる．ユーザ意図推定はユーザ発話を意図のカテゴリに分類する処理である．一方，スロットフィリングはユーザ発話の意味的な内容をスロットと値の組として出力する処理であり，例えば「ニューヨークからシカゴまで行きます」というユーザ発話に対し，出発地=“ニューヨーク”，目的地=“シカゴ” を出力する．この一般的な枠組みから言えば，我々の取り組む課題はスロットフィリングに相当する．

NLU におけるスロットフィリングは，発話中の各単語に対して，意味的なスロットの IOB タグ [9] を付与する系列ラベリング問題として定式化されることが多い．近年では，多くの文献で，この問題を解くための手法として Recurrent Neural Network（RNN）[2], [3], [10], [11], [12], [13], [14], [15] や Long Short-Term Memory（LSTM）[16], [17] が用いられている．しかし，これらの系列ラベリングによる手法で捉えられるのは発話中で明示的に言及された意味的なスロットのみであり，我々が対象としている非明示的条件を抽出できない．

タスク指向対話システムでのスロットフィリングにおいて，非明示的なスロットの抽出を対象とした研究はほとんど存在しない．文献 [18] は，映画検索をおこなうタスク指向対話システムにおいて，ユーザ発話から，ユーザの求めている映画のジャンルを推定する課題に取り組んでいる．彼らの目的は，ユーザ発話中でジャンルについて明示的に言及されていない場合においてもジャンルの推定をおこなうことである．例えば，彼らは “I wanna watch a movie that will make me laugh.” というユーザ発話から，ユーザが求める映画のジャンルが *comedy* であるということを推定する．彼らの研究の動機は我々とほとんど同じであるが，彼らはジャンルの推定結果だけを出力し，そのユーザ発話からそのジャンルが推定できる理由を提示していない点で，我々の目的とは異なる．また，彼らは映画のジャンルという 1 つの属性についてのみ推定をおこなっているが，我々は複数のデータベースフィールドに対する推定をおこなう．

近年では，パイプライン処理による伝統的なタスク指向対話システムとは異なり，ひとつひとつのモジュールを陽に作成せず，ユーザ発話から直接システム発話を生成する End-to-End のタスク指向対話システムも提案されている．文献 [19] は，知識ベースの検索を要求するユーザ発話に対

して，検索をおこない，その結果を提示することが可能な End-to-End の対話システムを提案している．彼らのシステムは NLU を陽にはおこなわない．このシステムは，与えられたユーザ発話に対し，その検索結果が得られた理由について考慮しておらず，システムの出力に対する説明をおこなわない．我々のシステムでは，非明示的条件が与えられた際，データベースへのクエリを作成するために推論が必要となる．システムが必ず正しい推論をおこなう保証は無いため，自然で効率的な対話を実現するためには，システムによる推論理由の説明やユーザへの確認を行う発話を用意することが必要である．しかし，現在の End-to-End の枠組みでは，それらを実現することは難しい．

3. データと問題設定

本論文では，対話コーパスとして不動産検索対話コーパス [20] を用いる．このコーパスは物件を探す客と不動産屋を演じる 2 名の作業者間で行われる日本語テキストチャット対話を収集している．対話の目的は客の物件に対する希望を不動産屋が聞き出すことである．不動産屋は実際にデータベースの検索をおこなうことはしないが，検索に必要な情報が十分得られたと判断した場合に対話を終了する．それぞれの対話において，客は 10 種類のプロファイルのうち 1 つが割り当てられ，そのプロファイルに合致する条件の物件を希望するよう指示されている．客のプロファイルは不動産屋には開示されない．実際のプロファイルのひとつの例として，「2 年ほど付き合った彼氏と同棲することになり，これまでのワンルームから引っ越すことになった女性．これを機に料理に力を入れたいため，コンロが多く使いやすいキッチンがある物件を希望している．」がある．

コーパス中の対話数は 986 対話，総発話数は 29,058 発話であり，そのうち不動産屋の発話が 14,571 発話，客の発話が 14,487 発話である．また，1 対話あたりの平均発話数は 29.5 発話である．

我々は，このコーパス中の各発話に対し，表 1 に示すデータベースフィールドタグをアノテーションした [6]．これらのデータベースフィールドタグは，日本の不動産情報サイト SUUMO[*1] で不動産検索をおこなう際に指定可能な検索条件をもとに設計された．また，これら 38 種類のタグに加え，どのデータベースフィールドタグにも該当しない発話に付与する【その他】タグを定義した．この【その他】タグが付与された発話に非明示的条件が含まれていることを期待している．さらに，アノテータが【その他】タグを付与する際には，同時に，その意味内容を自由記述するよう指示した．

本論文の課題は，【その他】が付与された客の発話を 38 種類のデータベースフィールドタグに分類し，同時に，その分類の根拠を表す客の発話中の表現を抽出することである．対象となる客の発話は，その直前の不動産屋側の発話と結合し，ひとつのテキストして扱う[*2]．これは，【その他】タグが付与された客の発話の中に単なる肯定のような発話が含まれるためである．例えば，不動産屋からの「奥さんとお住いになるということですか？」という発話に対する「はい，そうです.」という客の発話がこれに該当する．この場合，客の発話単体では何の情報も得られないが，直前の不動産屋の質問と組み合わせることにより「住む人数」という情報が得られ，【間取りタイプ】のようなデータベースフィールドへの分類が可能になる．このように，分類に必要な情報を集めるために，客の発話だけでなく直前の不動産屋の発話も含めてひとつのテキストとして扱う．本論文ではこの発話のまとまりを，発話チャンクと呼ぶ．発話チャンクは全部で 2,642 個作成された．

場所に関するタグ	部屋に関するタグ
- 【沿線】	- 【室内設備】
- 【駅徒歩】	- 【冷暖房】
- 【駅利便性】	- 【収納】
- 【エリア】	- 【バス・トイレ】
- 【周辺環境】	- 【キッチン】
- 【土地特徴】	- 【テレビ・通信】
- 【目的地からの時間】	物件の情報に関するタグ
建物に関するタグ	- 【物件種別】
- 【築年数】	- 【賃料】
- 【間取りタイプ】	- 【価格】
- 【専有面積】	- 【入居条件】
- 【部屋の位置】	- 【入居時期】
- 【一部屋の広さ】	- 【物件のターゲット】
- 【構造】	- 【販売状況】
- 【日当たり】	- 【提示可能情報】
- 【建物設備】	- 【所有権】
- 【セキュリティ】	- 【賃料の割引】
- 【建物の階数】	- 【補助金】
- 【戸数】	- 【証明書類】
- 【リフォーム・リノベーション】	- 【瑕疵保証】

表 1 データベースフィールドタグ

4. 提案手法

本論文では，サポートベクターマシン (SVM) による手法と Recurrent Convolutional Neural Network（RCNN）による手法の 2 つを提案する．

4.1 SVM による手法

それぞれのデータベースフィールドタグについて，入力

[*1] http://suumo.jp

[*2] 【その他】タグが付与された客の発話の直前の発話が【その他】タグが付与されていない客の発話であった場合にも，それらをひとつの塊とする．

の発話チャンクからそのデータベースフィールドが抽出できるか否かを分類する 2 値分類器を線形 SVM によって作成する．すなわち，全部で 38 個の 2 値分類器が作成される．入力の発話チャンクが与えられたとき，システムの最終的な出力は分類器が正と判断したデータベースフィールドのリストとなる．SVM の素性として，発話チャンクの bag-of-words を用いる．素性に使用する単語はコーパス中で 2 回以上出現する名詞，動詞，形容詞，副詞である．発話チャンクの形態素解析には，MeCab[*3]を使用する．また，数と固有名詞については，抽象化のためにそれぞれ NUM と PROP という記号に置き換える．素性ベクトルの次元数は 1,730 である．

分類の根拠抽出には，各データベースフィールドタグの分類器で学習された素性の重みを用いる．入力の発話チャンク中で，ある閾値以上の重みを持つ単語を全て抽出し，それらを分類の根拠とする．

4.2 RCNN による手法

我々は，文献 [7] によって提案された手法を拡張することによって，我々の課題に対する手法を提案する．彼らの手法は，商品レビューのテキストが入力として与えられたとき，それぞれの評価項目についてのユーザ評価を回帰によって推定し，また，その結果の根拠となる部分を入力テキストから抽出する．彼らのシステムは，回帰問題を解くニューラルネットワーク（エンコーダ）と根拠の抽出をおこなうニューラルネットワーク（ジェネレータ）の 2 つの要素からなる．エンコーダの学習はレビューテキストに対する真のユーザ評価を用いた教師あり学習によっておこなわれる．一方，ジェネレータの学習は教師なし学習でおこなわれる．彼らは，より短く，より連続した単語列が根拠として好ましいという仮定のもと，ジェネレータがそのような根拠抽出をおこなうよう損失関数を設計している．また，エンコーダは，ジェネレータが正しく根拠を抽出していると仮定し，ジェネレータによって抽出された単語のみをユーザ評価の推定に用いる．彼らは，教師あり学習でエンコーダの性能を向上させることで，間接的にジェネレータの性能を向上させることを狙っている．エンコーダとジェネレータのネットワークをそれぞれ図 2, 図 3 に示す．

彼らの手法を我々の課題に対して拡張するために，エンコーダの損失関数を変更する．彼らのシステムでは回帰のために二乗損失を用いていたが，我々は 2 値分類をおこなうため，代わりに交差エントロピーを使用する．$\tilde{y}_i = \mathrm{enc}_i(\boldsymbol{z}_i, \boldsymbol{x}) \in \{0, 1\}$ を i 番目のデータベースフィールドタグに対する 2 値分類の結果とすると，エンコーダの損失関数 $\mathcal{L}_i(\boldsymbol{z}_i, \boldsymbol{x}, y_i)$ は式 (1) のように定義される．

$$\mathcal{L}_i(\boldsymbol{z}_i, \boldsymbol{x}, y_i) = -(y_i \log(\tilde{y}_i) + (1 - y_i) \log(1 - \tilde{y}_i)). \quad (1)$$

[*3] http://taku910.github.io/mecab/

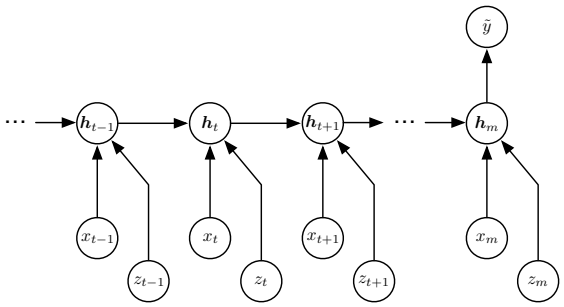

図 2 エンコーダ

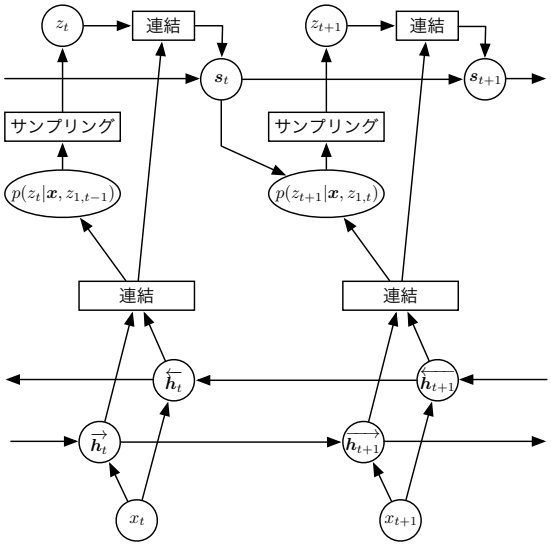

図 3 ジェネレータ

ここで，$\boldsymbol{x}$ は長さ m の入力単語列，$\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ は各単語がジェネレータによって根拠として選択されたかどうかを示す 2 値ベクトルを表し，$\mathrm{enc}_i(\boldsymbol{z}_i, \boldsymbol{x})$ は，$\boldsymbol{z}_i$ の要素が 1 である位置に対応する単語のみをエンコーダに入力して得られた結果を意味する．また，y_i は 2 値分類の正解を表す．ジェネレータの損失関数は，文献 [7] で使用されているものと同様であり，式 (2) で表される．

$$\Omega_i(\boldsymbol{z}_i) = \lambda_1 ||\boldsymbol{z}_i|| + \lambda_2 \sum_{t=1}^{m} |z_{it} - z_{i(t-1)}|. \quad (2)$$

ここで，$z_0 = 0$ とする．式 (2) の第 1 項は根拠が短くなることを，第 2 項は根拠が連続することを，それぞれ促す罰則項である．これら 2 つの罰則項のバランスを調整するために 2 つのハイパーパラメータ λ_1 と λ_2 がある．i 番目のデータベースフィールドに対するシステム全体の目的関数は，

$$\mathrm{cost}_i(\boldsymbol{z}_i, \boldsymbol{x}, y_i) = \mathcal{L}_i(\boldsymbol{z}_i, \boldsymbol{x}, y_i) + \Omega_i(\boldsymbol{z}_i) \quad (3)$$

となる．

5. 評価実験

5.1 データ

我々は，コーパス中の 986 対話を，客のプロファイルの

分布が変わらないように 10 分割し，そのうち 9 つを学習データ，残り 1 つをテストデータとして用いた．そして，3 節で説明したように発話チャンクを抽出した．その際，発話チャンク内のユーザ発話が，挨拶や対話の開始，終了のような対話管理レベルの発話であるような発話チャンクは除いた．学習データは 2,379 個，テストデータは 263 個の発話チャンクからなる．

それぞれの発話チャンク内のユーザ発話には，3 節で説明したように，【その他】タグとその意味内容の記述が付与されている．我々は，意味内容の記述をデータベースフィールドタグに写像することで，分類の正解を発話チャンクに付与した．例えば，「一人暮らしをしたいのですが.」という発話に付与された「住む人数」という意味内容の記述は，【間取りタイプ】と【専有面積】の 2 つのタグに写像される．このように複数のデータベースフィールドタグが 1 つの発話チャンクに付与されることもある．また，分類の正解として付与されたデータベースフィールドに対して，発話チャンク中の各単語がそのフィールドへの分類の根拠に含まれるか否かを付与し，これを根拠抽出の正解とした．これらのアノテーションは著者のうち 1 人がおこなった．

5.2 評価尺度

各データベースフィールドの 2 値分類の評価尺度として精度，再現率，F 値を使用する．また，分類の根拠抽出を評価するために，単語ごとの F 値，および BLEU [21] と ROUGE [22] を用いる．なお，根拠抽出の評価は，分類に正解した例に対してのみおこなう．$\tilde{\boldsymbol{z}} = (\tilde{z_1}, \tilde{z_2}, \cdots, \tilde{z_m})$ と $\boldsymbol{z} = (z_1, z_2, \cdots, z_m)$ を，それぞれ推定された根拠と正解の根拠を表す 2 値ベクトルとする．ここで，これらのベクトルの要素が 1 となることは，対応する位置の単語が根拠として選択されることを意味する．$\tilde{\boldsymbol{z}}$ と $\boldsymbol{z}$ が与えられたとき，単語ごとの F 値は式 (7) によって計算される．

$$\mathrm{TP} = \sum_{t=1}^{m} \tilde{z}_t z_t, \tag{4}$$

$$\mathrm{FP} = \sum_{t=1}^{m} \tilde{z}_t (1 - z_t), \tag{5}$$

$$\mathrm{FN} = \sum_{t=1}^{m} (1 - \tilde{z}_t) z_t, \tag{6}$$

$$\mathrm{F} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{7}$$

BLEU と ROUGE の計算のために，推定された根拠と正解の根拠に含まれる n-gram の集合をそれぞれ $\tilde{G_n}$，G_n とし，式 (8)，(9) によって定義する．

$$\tilde{G}_n = \left\{ \{j\}_{j=t}^{t+n-1} \middle| 1 \le t \le m-n+1 \wedge \prod_{j=t}^{t+n-1} \tilde{z}_j = 1 \right\}, \tag{8}$$

$$G_n = \left\{ \{j\}_{j=t}^{t+n-1} \middle| 1 \le t \le m-n+1 \wedge \prod_{j=t}^{t+n-1} z_j = 1 \right\}. \tag{9}$$

これらの n-gram の集合を用いて，BLEU と ROUGE をそれぞれ式 (10)，(11) のように計算する．今回は，uni-gram と bi-gram を用いるため，$n \le 2$ までの $\tilde{G_n}$ と G_n を用いる．

$$\mathrm{BLEU} = \left(\prod_{n=1}^{2} P_n \right)^{1/2}, \tag{10}$$

$$\mathrm{ROUGE} = \left(\prod_{n=1}^{2} Q_n \right)^{1/2}. \tag{11}$$

ここで，

$$P_n = \begin{cases} \dfrac{|\tilde{G}_n \cap G_n|}{|\tilde{G}_n|} & (|\tilde{G}_n| > 0 \wedge n = 1) \\ \dfrac{|\tilde{G}_n \cap G_n| + 1}{|\tilde{G}_n| + 1} & (\text{otherwise}) \end{cases}, \tag{12}$$

$$Q_n = \begin{cases} \dfrac{|\tilde{G}_n \cap G_n|}{|G_n|} & (|G_n| > 0 \wedge n = 1) \\ \dfrac{|\tilde{G}_n \cap G_n| + 1}{|G_n| + 1} & (\text{otherwise}) \end{cases}, \tag{13}$$

である．

式 (12)，(13) において，共通の n-gram が存在しない場合に全体の値が 0 となることを防ぐために，$n = 1$ の場合を除き，分母と分子に 1 を足す平滑化をおこなっている．また，BLEU には短すぎる出力に対する罰則として brevity penalty を導入することが一般的であるが，本論文での評価では，短すぎる出力に対しては ROUGE の値が小さくなるため，BLEU に対する罰則は導入していない．

5.3 実験設定

データ数の制約から，本論文では【周辺環境】，【間取りタイプ】，【専有面積】の 3 つのデータベースフィールドタグについてのみ評価をおこなう．これらは，最も多くの発話チャンクに付与された 3 つのタグである．学習データ中でそれぞれのタグが付与された発話チャンク数は，【周辺環境】について 1,033，【間取りタイプ】について 974，【専有面積】について 964 である．

RCNN による手法には，抽出する根拠の単語長と根拠の連続性のバランスを調整する 2 つのハイパーラパメータ λ_1，λ_2 が存在する．これらのハイパーパラメータを決定するために，学習データ中から【周辺環境】が付与された

発話チャンクをランダムに 200 個抽出し，開発データとした．そして，残りの学習データで RCNN による手法の学習をおこない，開発データによって評価をおこなった．その結果，開発データにおいて，$\lambda_1 = 0.021$，$\lambda_2 = 0.003$ のときに最も良い性能を示すことがわかった．これらのハイパーパラメータを 3 つのタグについての評価実験で使用した．また，開発データで学習されたネットワークのパラメータを，評価実験における学習時の初期パラメータとして用いた．単語埋め込みベクトルとして，学習済みの日本語 Wikipedia エンティティベクトル[*4]を用いた．各単語ベクトルの次元数は 200 である．

SVM による手法にも，根拠単語の抽出に使用する閾値がハイパーパラメータとして存在する．この手法では，その閾値よりも大きな重みを持つ単語を根拠として抽出する．この閾値を決定するために RCNN による手法で使用したものと同じ開発データを用い，同様の方法で最適な閾値を求めた．単語の重みは，式 (14) によって正規化される．結果的に求められた閾値は 0.58 であった．

$$\hat{w} = \frac{w - w_{\min}}{w_{\max} - w_{\min}} \tag{14}$$

5.4 実験結果と考察

表 2 と表 3 はそれぞれ，データベースフィールドへの分類と，根拠抽出の評価結果を示す．表 3 には，比較のために，SVM による手法において，発話チャンク中で最も重みの大きい 1 単語を根拠として抽出するように変更した手法（SVM_1）の評価結果も含めている．両方の課題において，SVM による手法が RCNN による手法よりもより良好な結果を示した．

データベースフィールドへの分類

表 2 は，全ての評価尺度において，SVM による手法が RCNN による手法に勝っていることを示している．RCNN による手法は，データベースフィールドへの分類を学習する際，発話チャンク全体ではなく，分類の根拠として抽出された単語のみを使用している．これは，データベースフィールドへの分類の性能が，分類の根拠抽出の性能に大きく依存することを意味する．表 3 に示すように，分類の根拠抽出の単語ごとの F 値は 0.5 未満である．このことは，RCNN による手法が，誤った根拠という不十分な情報のみを用いてデータベースフィールドへの分類をおこなわなければならないことを示している．

分類の性能と根拠抽出の結果との間の関係をより詳しく調べるために，データベースフィールドへの分類結果のうち，真陽性である例と偽陰性である例の数をそれぞれ数えた．表 4 は，推定の根拠と正解の根拠との間に少なくとも 1 単語の重なりがあるかどうかによって，真陽性と偽陰性の例をそれぞれ分割した結果を示す．「有り」の行は，少なくとも 1 単語が，推定と正解の根拠とで重複している例の数を示し，「無し」の行は，1 つも単語の重なりのない例の数を表す．例えば，【周辺環境】において，重なりが「有り」である 90 例のうち 73 例（81%）は分類結果が正しい（真陽性である）のに対し，重なりが「無し」であり分類結果が正しい例は 23 例中 14 例（60%）にとどまっている．この傾向は他の 2 つのデータベースフィールドタグについても同様である．このことから，根拠抽出の性能がデータベースフィールドへの分類性能に大きな影響をもたらすことがわかる．

しかし，表 4 は同時に，RCNN による手法の根拠抽出が完全に誤っているにも関わらず，分類に成功している例が存在することも示している．我々は【周辺環境】タグについて，推定の根拠と正解の根拠の重なりがないものの，分類に正解している 14 例（表 4 中の，重なりが「無し」で真陽性である例）について分析をおこなった．その結果，分類の手がかりとなっているものの，根拠の正解には含まれない単語が存在することがわかった．例えば，14 例中 3 例で根拠として抽出されていた「近く」という単語は，学習データ中の正例において，負例よりも 2 倍多く出現する．この単語は，正解において分類の根拠には含まれていないものの，【周辺環境】タグへの分類の手がかりと考えることができる．データベースフィールドへの分類性能を向上させるためには，このような，根拠の正解としてアノテーションされていないものの，分類に役立つ情報を利用できるよう手法を拡張する必要がある．

我々はさらに，【周辺環境】への分類において誤った例について個々に調査した．その結果，RCNN による手法において偽陽性であった 25 例中 10 例，および SVM による手法において偽陽性であった 24 例中 11 例について，アノテーションの不一致があることを発見した．例えば，「場所の希望はありますか？—-郊外を希望します。」という発話チャンクは負例とアノテーションされている一方，ほぼ同様の内容である「どのような物件をお探しですか？—-将来のことを考えて，少し郊外へ越そうと思うのですが。」という発話チャンクは，正例としてアノテーションされている．我々は，発話チャンクに対してデータベースフィールドタグを正解としてアノテーションする際，発話そのものを参照せず，【その他】タグが付与された発話に同様に付与されている内容の記述のみからデータベースフィールドタグへの写像をおこなった．上記の例では，前者の発話チャンクには「大まかなエリア」という内容が付与されているが，後者には「住環境」という内容が付与されている．元々の発話の内容はおおよそ同じであるにも関わらず，内容の記述が異なったことにより，このようなアノテーションの不一致が引き起こされた．より安定して一貫したアノテーションが可能となるように，より頑健で具体的なガイ

[*4] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

	【周辺環境】			【間取りタイプ】			【専有面積】		
手法	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
SVM	**0.789**	**0.796**	**0.793**	**0.918**	**0.865**	**0.891**	**0.891**	**0.874**	**0.882**
RCNN	0.777	0.770	0.773	0.881	0.856	0.868	0.873	0.864	0.868

表 2 データベースフィールドへの分類の評価結果

	【周辺環境】			【間取りタイプ】			【専有面積】		
手法	単語ごとの F 値	BLEU	ROUGE	単語ごとの F 値	BLEU	ROUGE	単語ごとの F 値	BLEU	ROUGE
SVM_1	0.514	0.768	0.467	0.440	**0.808**	0.364	0.420	0.763	0.345
SVM	**0.530**	**0.787**	0.552	**0.533**	0.802	0.467	**0.507**	**0.773**	0.462
RCNN	0.458	0.534	**0.576**	0.452	0.625	**0.475**	0.436	0.521	**0.485**

表 3 根拠抽出の評価結果

ドラインを作成することは今後の課題である．

根拠抽出

表 3 は，全てのデータベースフィールドの ROUGE を除いた評価尺度において SVM による手法が RCNN による手法よりも良好な結果を示したことを示す．

表 5 はそれぞれの手法での推定の根拠と正解の根拠において抽出された単語数と区間数を示す．ここで，区間とは，連続して抽出された単語列のことを指す．

3 つのデータベースフィールドについて，どちらの手法も正解よりも少ない単語を根拠として抽出している一方，抽出された区間数は正解よりも多い．その結果，1 区間あたりの単語数（単語数/区間数）は，両手法で正解の半分未満となっている．SVM による手法では，1 区間あたりの単語数がほぼ 1 である．一方，RCNN による手法の推定結果では，1 区間あたりの単語数は SVM による手法のそれよりも若干多く，より長い根拠を抽出していることがわかる．表 3 で RCNN による手法が ROUGE において SVM による手法を上回ったのはこれが要因である．また，このことは，より連続した根拠を選ぶように設計したジェネレータの損失関数が期待通りのはたらきをしていることを示している．しかし，推定された区間の長さは正解と比較すると依然短い．

表 3 中の SVM_1 と SVM の結果を比較すると，ほぼ全ての BLEU と ROUGE において SVM のほうが SVM_1 よりも高い性能を達成している．素性の重みが閾値以上である単語を全て根拠として抽出する SVM のほうが ROUGE が高くなることは必然であるが，BLEU も 3 つのタグのうち 2 つにおいて高いことから，閾値以上の重みを持つ単語には根拠となる単語が多いことがわかる．

根拠抽出は，発話チャンク中の各単語に対する，根拠に含まれるか否かの 2 値分類として定式化することができる．表 6 は【周辺環境】タグにおける混同行列である．この行列は，両手法において偽陰性の誤りのほうが偽陽性の誤りよりも多いことを示している．偽陰性の誤りのうち，最も多い単語は「が」や「に」のような助詞であった．言い換えれば，両手法はこれらの助詞を飛ばし，名詞や動詞のような内容語を抽出する傾向にある．これらの助詞はどのようなデータベースフィールドにおいても出現するため分類に重要な素性ではない．しかし，対話システムが，後の処理として，データベースフィールドへの分類に対する確認発話を生成する際には，これらの助詞も分類の根拠に含まれることが望ましい．

RCNN による手法によって出力された【周辺環境】タグに対する根拠について，個々に分析をおこなったところ，この手法は，全てのテストケースにおいてクエスチョンマーク（「？」）の直後の単語を根拠として抽出することがわかった．このクエスチョンマークは，ほとんどの場合において，不動産屋の質問の最後の単語である．【周辺環境】タグへの分類の強力な手がかりとなる語として「治安」という単語があるが，この単語は学習データにおいて，41.3%が，クエスチョンマークの直後に出現する．RCNN による手法のジェネレータはこの共起を学習したと考えられる．しかし，実際のテストデータにおいて，クエスチョンマークの直後の単語が根拠の正解に含まれる例は全体の 40.4%しかなく，これによって根拠抽出の性能が低下している..

RCNN による手法はデータベースフィールドタグへの分類とその根拠抽出の両方において，SVM による手法よりも性能が劣っている．本論文では，学習データに 2,379 個の発話チャンクのみを使用したが，RCNN による手法の元となる論文 [7] では，約 8 万から 9 万のレビューテキストを用いており，RCNN による手法でより高い性能を達成するためにはより多くのデータが必要であることが推測される．

6. Conclusion

本論文は，データベース検索を行うタスク指向型対話を

	【周辺環境】		【間取りタイプ】		【専有面積】	
重なり	真陽性	偽陰性	真陽性	偽陰性	真陽性	偽陰性
有り	73	17	86	10	75	11
無し	14	9	3	5	14	3

表 4 推定された根拠と正解の根拠との間に単語の重なりがあるかに基づいて真陽性の例と偽陽性の例を分割した結果

	【周辺環境】			【間取りタイプ】			【専有面積】		
	単語数	区間数	単語数/区間数	単語数	区間数	単語数/区間数	単語数	区間数	単語数/区間数
SVM	1.60	1.54	1.04	2.48	2.33	1.07	2.49	2.32	1.07
RCNN	3.21	2.39	1.34	2.97	2.06	1.44	3.57	2.12	1.69
正解	3.42	1.23	2.78	4.58	1.20	3.81	4.58	1.20	3.81

表 5 根拠区間の大きさ

			正解	
			正	負
推定	正	SVM	102	64
		RCNN	92	144
	負	SVM	193	2,353
		RCNN	165	2,038

表 6 根拠抽出の混同行列

対象として，ユーザ発話中で明示的に述べられていないユーザ要求の解釈をおこなう手法を提案した．我々はこのように明示的に述べられないユーザ要求を非明示的条件と呼び，その解釈を，ユーザ発話を関連するデータベースフィールドに分類し，また同時にその根拠となるユーザ発話中の文字列を抽出する課題として定式化した．この課題に対する 2 つの手法として，サポートベクターマシンに基づく手法と，分類と根拠となる部分文字列の抽出を同時に行うニューラルネットワークによる手法を実装した．不動産に関する対話のコーパスを利用した評価実験の結果，サポートベクターマシンに基づく手法がより良好な結果を示すことがわかった．この結果は，使用したコーパスが小さく，ニューラルネットワークによる手法を学習するのに十分でなかったことを示唆している．また，エラー分析を通して，いくつかの誤りは，アノテーションの不一致によって引き起こされていることがわかった．今後の課題として，より安定して一貫したアノテーションが可能となるように，より頑健で具体的なガイドラインを作成する必要がある．

本論文では，ユーザ発話をデータベースフィールドに分類することと，発話からその根拠を抽出することにのみ焦点を当てていたが，実際にユーザ発話からデータベースクエリを生成するためには，データベースフィールドの値も抽出することが必要となる．この問題に取り組むことは今後の課題である．

参考文献

[1] Raymond, C. and Riccardi, G.: Generative and discriminative algorithms for spoken language understanding, *Eighth Annual Conference of the International Speech Communication Association*, pp. 1605–1608 (2007).

[2] Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-tur, D., He, X., Heck, L., Tur, G., Yu, D. and Zweig, G.: Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 3, pp. 530–539 (2015).

[3] Liu, B. and Lane, I.: Joint Online Spoken Language Understanding and Language Modeling with Recurrent Neural Networks, *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2016)*, pp. 22–30 (2016).

[4] Hemphill, C. T., Godfrey, J. J. and Doddington, G. R.: The ATIS Spoken Language Systems Pilot Corpus, *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pp. 96–101 (1990).

[5] Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A. and Shriberg, E.: Expanding the Scope of the ATIS Task: The ATIS-3 Corpus, *Proceedings of the Workshop on Human Language Technology*, HLT '94, pp. 43–48 (1994).

[6] Fukunaga, S., Nishikawa, H., Tokunaga, T., Yokono, H. and Takahashi, T.: Analysis of Implicit Conditions in Database Search Dialogues, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2741–2745 (2018).

[7] Lei, T., Barzilay, R. and Jaakkola, T.: Rationalizing Neural Predictions, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 107–117 (2016).

[8] Chen, H., Liu, X., Yin, D. and Tang, J.: A Survey on Dialogue Systems: Recent Advances and New Frontiers, *SIGKDD Explor. Newsl.*, Vol. 19, No. 2, pp. 25–35 (online), DOI: 10.1145/3166054.3166058 (2017).

[9] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Third Workshop on Very Large Corpora* (1995).

[10] Mesnil, G., He, X., Deng, L. and Bengio, Y.: Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, *INTERSPEECH-2013*, pp. 3771–3775 (2013).

[11] Yao, K., Zweig, G., Hwang, M.-y., Shi, Y. and Yu, D.: Recurrent Neural Networks for Language Understanding, *INTERSPEECH-2013*, pp. 2524–2528 (2013).
[12] Vu, N. T., Gupta, P., Adel, H. and Schutze, H.: BI-DIRECTIONAL RECURRENT NEURAL NETWORK WITH RANKING LOSS FOR SPOKEN LANGUAGE UNDERSTANDING, *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 6060–6064 (2016).
[13] Jaech, A., Heck, L. and Ostendorf, M.: Domain Adaptation of Recurrent Neural Networks for Natural Language Understanding, *INTERSPEECH-2016*, pp. 690–694 (2016).
[14] Liu, B. and Lane, I.: Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling, *INTERSPEECH-2016*, pp. 685–689 (2016).
[15] Bapna, A., Tur, G., Hakkani-tur, D. and Heck, L.: Sequential Dialogue Context Modeling for Spoken Language Understanding, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2017)*, pp. 103–114 (2017).
[16] Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G. and Shi, Y.: SPOKEN LANGUAGE UNDERSTANDING USING LONG SHORT-TERM MEMORY NEURAL NETWORKS, *Spoken Language Technology Workshop (SLT)*, pp. 189–194 (2014).
[17] Hakkani-t, D., Tur, G., Celikyilmaz, A., Chen, Y.-n., Gao, J., Deng, L. and Wang, Y.-y.: Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM, *INTERSPEECH-2016*, pp. 715–719 (2016).
[18] Celikyilmaz, A., Hakkani-tür, D. and Tur, G.: STATISTICAL SEMANTIC INTERPRETATION MODELING FOR SPOKEN LANGUAGE UNDERSTANDING WITH ENRICHED SEMANTIC FEATURES, *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pp. 216–221 (2012).
[19] Eric, M., Krishnan, L., Charette, F. and Manning, C. D.: Key-Value Retrieval Networks for Task-Oriented Dialogue, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2017)*, pp. 37–49 (2017).
[20] Takahashi, T. and Yokono, H.: Two persons dialogue corpus made by multiple crowd-workers, *Proceedings of the 8th International Workshop on Spoken Dialogue Systems (IWSDS 2017)* (2017). 6 pages.
[21] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318 (2002).
[22] Lin, C.-Y. and Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 71–78 (2003).