Sequence-to-Sequence モデルを用いた 非明示的条件の分類とその根拠抽出

福永 隼也† 西川 仁† 徳永 健伸† 横野 光‡ 高橋 哲朗‡
†東京工業大学 情報理工学院 ‡株式会社富士通研究所
fukunaga.s.ab@m.titech.ac.jp {hitoshi,take}@c.titech.ac.jp
{yokono.hikaru,takahashi.tet}@fujitsu.com

1 はじめに

データベース（DB）検索を行うタスク指向型対話システムは，DB 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 DB フィールドとその値を抽出する必要がある．DB 検索対話において，ユーザ発話中に明示的に出現する検索条件を抽出する研究はこれまで Liu and Lane (2016) など，多く行われてきた．

一方，実際の対話には，DB 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し，対話システムがそのような情報を利用することで，より自然で効率的な DB 検索を行うことが可能になる．例として，不動産業者と不動産を探す客の対話を考える．不動産検索において，客の家族構成は物件の広さを絞り込む上で有用な情報であるが，家族構成は物件の属性ではなく客の属性であるため，通常，不動産 DB には含まれない．客の家族構成のように，DB フィールドには直接対応しないが，DB 検索を行う上で有用な情報を**非明示的条件**と呼ぶ (Fukunaga et al., 2018b)．

Fukunaga et al. (2018a) は，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，以下の 2 つのサブタスクを同時に行う新しいタスクを提案し，そのための手法として，サポートベクタマシン（SVM）と Recurrent Convolutional Neural Network（RCNN）(Lei et al., 2016) による手法を実装した．

(1) 非明示的条件を含むユーザ発話を，DB フィールドとその値の組（検索条件）へ変換する．

(2) ユーザ発話中から，(1) で行った検索条件への変換の根拠となる部分を抽出する．

サブタスク (1) は，非明示的条件を含む発話から DB へのクエリを作成するために必要な処理である．図 1 に示す対話では，客の発話に含まれる「一人暮らし」という文言から，〈間取り ≤1LDK〉という検索条件へ変換できる．Fukunaga et al. (2018a) はサブタスク (1) の第一段階として DB フィールドの特定に取り組み，1 つのユーザ発話が複数の DB フィールドに関連しうることから，この問題をマルチラベル分類問題として定式化した．

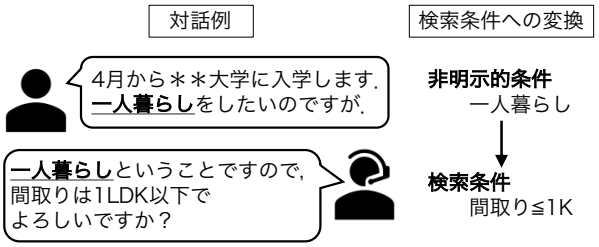

図 1: 対話と非明示的条件から検索条件への変換の例

サブタスク (2) によって抽出した根拠はユーザがシステムの判断に納得するために有用である．非明示的条件から検索条件への変換には例外も存在するため，システムの解釈が正しいかどうかをユーザに確認する場合がある．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.」というユーザ発話をシステムが〈間取り ≤1LDK〉という検索条件へ変換したとする．このとき，単に「間取りは 1LDK 以下でよろしいですか？」と確認するよりも，「一人暮らしということですので,」とシステムが判断した根拠を追加することで，ユーザはその検索条件が得られた理由を理解でき，また，より自然な対話となる．

本稿では，Hori et al. (2016) による Attention 機構付き Sequence-to-Sequence（seq2seq）モデルを根拠抽出が行えるように拡張し，上記の課題に適用する．評価実験の結果，SVM や RCNN による手法よりも良好な性能を示した．

2 Attention 機構付き seq2seq モデル

対話データのマルチラベル分類手法として Hori et al. (2016) が提案する Attention 機構付き seq2seq モデルによる手法を根拠抽出を行うよう拡張し，我々の課題に適用する．この手法は，Dialog State Tracking Challenge 5 (Kim et al., 2016) のために提案された手法であり，入力として対話中の単語列を受け取ってスロットと値の組の系列を出力する．本稿では，出力を DB フィールドタグの系列とする．

モデルは，エンコーダとデコーダからなる．$\boldsymbol{x}_t (1 \leq t \leq m)$ を入力発話の t 番目の単語の埋め込みベクトルとすると，エンコーダでは以下の双方向 LSTM を用いて，各単語に対応する隠れ状態ベクトル $\boldsymbol{h}_t$ を計算する．

$$\boldsymbol{h}_t = [\boldsymbol{h}_t^{(f)}; \boldsymbol{h}_t^{(b)}], \quad (1)$$

$$\boldsymbol{h}_t^{(f)} = LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^{(f)}), \quad (2)$$

$$\boldsymbol{h}_t^{(b)} = LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t+1}^{(f)}). \quad (3)$$

デコーダでは，Attention 機構によって重み付けした各単語の隠れ状態ベクトルを用いて，DB フィールドタグの系列を出力する．i 番目の出力における t 番目の単語の隠れ状態ベクトルに対する重み $\alpha_{i,t}$ を，

$$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_{t=1}^{m} \exp(e_{i,t})}, \quad (4)$$

$$e_{i,t} = \boldsymbol{w}^T \tanh(W\boldsymbol{s}_{i-1} + V\boldsymbol{h}_t + \boldsymbol{b}). \quad (5)$$

によって計算する．エンコーダの隠れ状態ベクトルの重み付き平均 $\boldsymbol{g}_i = \sum_{t=1}^{m} \alpha_{i,t}\boldsymbol{h}_t$ を LSTM に入力し，各 DB フィールドの確率を表すベクトル $\boldsymbol{p}_i$ を出力する．

$$\boldsymbol{s}_i = LSTM(\boldsymbol{s}_{i-1}, \boldsymbol{y}_i, \boldsymbol{g}_i), \quad (6)$$

$$\boldsymbol{p}_i = \mathrm{softmax}(W_{SO}\boldsymbol{s}_{i-1} + W_{GO}\boldsymbol{g}_i + \boldsymbol{b}_{SO}). \quad (7)$$

ここで，$\boldsymbol{y}_i$ は，i 番目の出力を表す one-hot ベクトルであり，学習時には正解を与え，テスト時には $\boldsymbol{p}_i$ において最も確率の高い DB フィールドを与える．本来，各発話に付与される DB フィールドタグは系列では無いが，学習データとして与える際には，対応する添字の小さい順に正解のベクトルを作成し系列とする．損失関数には交差エントロピー損失を用いる．

Hori et al. (2016) の手法の拡張として，我々は Attention 機構による重み $\alpha_{i,t}$ が閾値よりも大きい単語を，i 番目に出力される DB フィールドの根拠として抽出する．閾値には，定数か，各発話に含まれる単語数の逆数を用いる．単語数の逆数を閾値に用いる理由は，発話中の単語数の違いを考慮するためである．Attention 機構による重みの総和は 1 となるため，仮に発話中の全単語に均等な重みを割り当てた場合，単語数が多いほど各単語に割り当てられる重みは小さくなる．したがって，閾値を定数とすると，発話中の単語数によって閾値の価値が変わってしまうため，これを防ぐために閾値として単語数の逆数を用いる．

3 データセット

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 (Takahashi and Yokono, 2017; Fukunaga et al., 2018b) を用いる．このコーパスは物件を探す客と不動産業者を演じる 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たものである．対話の目的は客の物件に対する希望を不動産業者が聞き出すことである．不動産業者は実際に DB 検索を行わず，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．客のプロファイルは不動産業者には開示されない．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，そのうち不動産業者の発話が 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．

コーパス中の各発話には，どの DB フィールドに該当するかを表す【周辺環境】や【間取りタイプ】など 38 種類の DB フィールドタグが付与されている．[1] また，いずれの DB フィールドにも該当しない発話に対しては【その他】タグが付与されている．本稿では，【その他】が付与された客の発話に非明示的条件が含まれると仮定し，それらを 38 種類の DB フィールドタグに分類することと，その分類の根拠となる発話の断片を抽出することを課題とする．客の発話は，分類のための情報を増やすために，その直前の不動産業者の発話と合わせてひとつのテキストとして扱う[2]．本稿では，これを発話チャンクと呼ぶ．

コーパス中の 986 対話を，客のプロファイルの分布を保ちながら 10 分割し，そのうち 9 つを学習データ，

[1] 以下，DB フィールドタグは【】で囲んで表記する．

[2] 客の発話が連続する場合，それらと直前の不動産業者の発話をひとつの塊とする．

残り 1 つをテストデータとして用いる．【その他】タグが付与された客の発話を含む発話チャンクを抽出した結果，学習データでは 2,379 個，テストデータでは 263 個の発話チャンクが収集できた．また，分類の正解となる DB フィールドタグを，学習データとテストデータに対して付与した．さらに，テストデータに対して，正解として付与された各 DB フィールドタグの根拠をアノテーションした．これらのアノテーションは，著者のうち 1 人が行った．

4 評価実験

4.1 実験設定

DB フィールドタグへのマルチラベル分類の評価は，各 DB フィールドタグへの二値分類としてみなし，その F 値を用いる．また，根拠抽出の評価には，4-gram までの F 値を用いる．ここで，根拠の n-gram とは，連続して根拠として抽出した n 単語を示す．

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】,【一部屋の広さ】,【エリア】の 5 つの DB フィールドタグについてのみ評価を行う．これらは，最も多くの発話チャンクに付与された 5 つのタグである．

手法の比較のために，Fukunaga et al. (2018a) で実装した SVM と RCNN による手法の評価実験も同様に行う．

各手法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 個抽出して開発データとして用いる．残りの学習データによって各手法をハイパーパラメータを変化させながら学習し，開発データで評価した結果，最も性能が良かったハイパーパラメータを評価実験に用いる．その結果，Attention 機構付き seq2seq モデルで根拠抽出に用いる重みの閾値は 0.0525 となった．

学習済みの日本語 Wikipedia エンティティベクトル[3]を固定の単語埋め込みベクトルとして用いる．各単語ベクトルの次元数は 200 である．

4.2 実験結果

各 DB フィールドタグについての分類結果と根拠抽出結果を表 1 と表 2 に示す．ここで，表中の「seq2seq」は Attention 機構付き seq2seq モデルによる手法の結果を表し，「定数」と「逆数」の行はそれぞれ，根拠抽出の閾値を定数とした場合と，発話チャンク中の単語数の逆数とした場合の結果である．各評価において，手法間で最大の値には下線を付している．DB フィールドへの分類については全体的に seq2seq モデルによる手法が良好な結果を示した．なお，この手法では，根拠抽出の閾値の設定によらず分類結果は同じであるため，表 1 では結果をまとめて記載している．根拠抽出については，$n \geq 2$ の場合に，seq2seq モデルによる手法，特に根拠抽出の閾値を定数とした手法の評価が高かった．

4.3 考察

DB フィールドタグへの分類については全体的に seq2seq モデルが，RCNN による手法よりも良い性能を示した．RCNN による手法は，根拠として抽出した単語のみを分類に用いる Hard Attention 機構を用いており，根拠として抽出できなかった単語の情報を分類に利用できない．それに対し seq2seq モデルは Soft Attention 機構を用いているため，入力中の全ての単語の情報を利用することができることが，より良い性能を示した要因であると考えられる．RCNN による手法の元論文である Lei et al. (2016) は，Attention 機構を用いたモデルよりも RCNN がより高い性能を示したと報告しているが，我々のデータセットのサイズは Lei et al. (2016) よりも小さく，うまく学習ができなかったと考えられる．

根拠抽出の閾値の設定方法による出力の違いを図 2 に示す．図から，閾値を定数とした場合では，正解の根拠と比較してより多くの単語を根拠として抽出することがわかる．1 発話チャンクあたり根拠として抽出された平均単語数を計算すると，正解の根拠が平均 5.08 単語であるのに対し，定数を閾値とした場合は平均 9.30 単語であり，正解と比較して 2 倍近くの単語を抽出してしまっていた．これは，今回閾値として用いた値が 0.0525 と非常に小さかったためである．一方，単語数の逆数を閾値とした場合においても，「を」や「か？」などの根拠とならない単語が単独で抽出されていた．これはこれらの単語に付与された Attention 機構の重みが比較的大きいことを示す．したがって，このような単語が単独で大きな重みを持たないように Attention 機構を工夫することや，重みを用いた根拠抽出手法の工夫を行うことが必要である．

[3] 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】	【一部屋の広さ】	【エリア】
SVM	0.793	0.891	0.882	0.872	0.854
RCNN	0.745	0.870	0.901	0.880	0.885
seq2seq	0.830	0.889	0.883	0.889	0.932

表 1: 各 DB フィールドタグへの二値分類の F 値

	【周辺環境】				【間取りタイプ】				【専有面積】			
n	1	2	3	4	1	2	3	4	1	2	3	4
SVM	0.353	0.033	0.000	0.000	0.502	0.076	0.000	0.000	0.486	0.076	0.000	0.000
RCNN	0.331	0.112	0.028	0.016	0.446	0.102	0.007	0.000	0.517	0.185	0.042	0.021
seq2seq												
定数	0.369	0.300	0.229	0.138	0.471	0.334	0.210	0.126	0.399	0.323	0.244	0.182
逆数	0.402	0.294	0.175	0.179	0.426	0.175	0.041	0.020	0.482	0.340	0.214	0.138

	【一部屋の広さ】				【エリア】			
n	1	2	3	4	1	2	3	4
SVM	0.493	0.104	0.000	0.000	0.350	0.000	0.000	0.000
RCNN	0.489	0.245	0.078	0.012	0.372	0.145	0.014	0.000
seq2seq								
定数	0.470	0.340	0.212	0.118	0.461	0.285	0.211	0.156
逆数	0.418	0.137	0.046	0.012	0.387	0.137	0.036	0.020

表 2: 各 DB フィールドタグにおいて抽出した根拠の n-gram の F 値

店 他に何かこだわり条件はございますか？
客 治安が良い地域を希望します。

図 2: seq2seq モデルの閾値の違いによる根拠抽出の比較．網掛け部は正解の根拠，破線と実線はそれぞれ閾値を定数と単語数の逆数とした場合の出力．

5 結論

本稿では，DB 検索対話において，非明示的条件を含むユーザ発話を DB フィールドへ変換し，同時にその根拠を抽出する課題に対し，Attention 機構付き seq2seq モデルを用いた手法を実装した．不動産検索対話コーパスを利用した評価実験の結果，この手法が先行研究の手法と比較し良好な結果を示すことがわかった．今回は根拠抽出手法として，Attention 機構による重みが閾値以上であれば根拠として抽出するという単純な手法を用いたが，根拠とならない単語を抽出してしまうという問題があった．根拠抽出手法を工夫することが今後の課題である．

参考文献

Shunya Fukunaga, Hitoshi Nishikawa, Takenobu Tokunaga, Hikaru Yokono, and Tetsuro Takahashi. Interpretation of implicit conditions in database search dialogues. In *Proceedings of COLING 2018*, pages 477–486, 2018a.

Shunya Fukunaga, Hitoshi Nishikawa, Takenobu Tokunaga, Hikaru Yokono, and Tetsuro Takahashi. Analysis of implicit conditions in database search dialogues. In *Proceedings of LREC 2018*, pages 2741–2745, 2018b.

T. Hori, H. Wang, C. Hori, S. Watanabe, B. Harsham, J. L. Roux, J. R. Hershey, Y. Koji, Y. Jing, Z. Zhu, and T. Aikawa. Dialog state tracking with attention-based sequence-to-sequence learning. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 552–558, 2016.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. The fifth dialog state tracking challenge. In *Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of EMNLP 2016*, pages 107–117, 2016.

Bing Liu and Ian Lane. Joint Online Spoken Language Understanding and Language Modeling with Recurrent Neural Networks. In *Proceedings of SIGDIAL 2016*, pages 22–30, 2016.

Tetsuro Takahashi and Hikaru Yokono. Two persons dialogue corpus made by multiple crowd-workers. In *Proceedings of IWSDS 2017*, 2017. 6 pages.