知識グラフによる特徴量ベクトルを用いたニューラルネットワークによる対話状態推定

村瀬　行俊[1], 吉野　幸一郎[1,2], 中村　哲[1]
[1]奈良先端科学技術大学院大学情報科学研究科
[2]科学技術振興機構　さきがけ
{y-murase, koichiro, s-nakamura}@is.naist.jp

1　はじめに

タスク指向型対話システムのモジュールである対話状態推定の精度を向上するため，The Dialog State Tracking Challenges (DSTCs) というシェアードタスクが過去５年に渡り開催されている [14, 13]．これまでの研究では，特に機械学習による識別的アプローチが高い精度を達成している [3]．DSTC5 では畳み込みニューラルネットワーク（CNN）による推定器の精度が最も高かった [11]．この推定器では単語埋め込み表現である Word2Vec [9] を入力特徴量として用いている．ただし，推定する状態の数に対して学習に使用できるラベル付きデータの量が限られているため，精度の向上が難しいという問題がある．また，入力となる発話に含まれる情報は限られることから，単純な特徴ベクトルの作り方では予測に必要な特徴を得ることが難しい．

このようにデータが限られている場合，外部の知識ベースをシステムの知識として用いることで，情報の欠落を補うことができる．近年では Web 上で不特定多数が編集可能な知識ベースを構築するプロジェクトが存在し，このような知識ベースを利用することが可能である [2, 12]．対話システムでも，こうした外部の知識情報として Web で収集した知識ベースが用いられている [8]．このように知識を用いることで元になる入力特徴から関連のある情報を推論し，付加して用いることが可能になる．

本研究では，外部の知識ベースを用いて入力発話からより多くの情報を抽出するために，知識グラフ上での推論による特徴量ベクトル構築を提案する．この特徴量ベクトルを全結合ニューラルネットワーク（FCNN）の入力として対話状態推定に用いた結果，既存研究の最も良い手法である CNN ベースの推定器に近い精度が確認された．また，提案手法を CNN ベースの推定器とアンサンブルすることで，Dialog State Tracking Challenge 4 (DSTC4)) [5] において最高精度を実現した．

2　対話状態推定

2.1　DSTC4

本研究では人間同士の旅行対話を扱っている DSTC4 での対話状態推定を行った．DSTC4 では副対話ごとにトピック（５種類）が与えられており，この副対話ごとに対話状態推定をする．DSTC4 のコーパスは３人のガイドと３５人の旅行者の対話をスカイプによって収録したものである．それぞれの対話は人手により書き起こしがされ，副対話ごとにアノテーションが付与されている．コーパスのサイズは対話数が 35 対話，発話数が 20,641 発話である．対話状態のアノテーションは副対話内の会話で取り上げられている内容を Slot-Value の組で付与されている．例えば，トピックが”Accommodation” であるとき，宿泊施設の種類である Slot の”TYPE” に対して”Hotel” や”Hostel” などの Value を取りうる．この Slot-Value の組み合わせは 5,608 組ある．DSTC4 では副対話の書き起こし済みの各発話に対して状態推定を行う．この際，対話履歴を用いてもよい．

2.2　CNN による対話状態推定

タスク指向型対話における対話状態推定タスクを扱った最新のチャレンジである DSTC5 で最も精度の高い識別器が CNN である．図 1 のように，CNN では発話を 1 つの行列として入力する．各行は各単語の固定長の単語埋め込みベクトルであり，発話内で観測された単語順で構成されている．これまでに提案されてきた発話内の語順を考慮したモデル [15] と同様に，このモデルでは密な単語ベクトル表現により語順に由来する意味的特徴量の利用を可能にしている [6]．

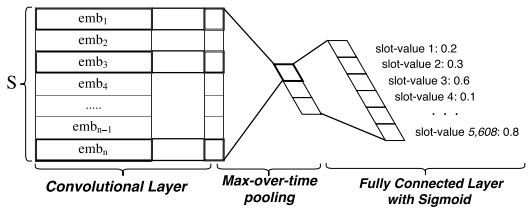

図 1: CNN モデルの概要図.

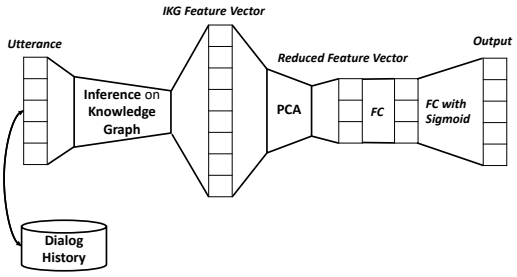

図 2: IKG-FCNN モデルの概要図.

2.3 外部知識上の推論による対話状態推定

DBpedia [1] や Wikidata [12] のような大規模な知識ベースが Web 上で構築され，日々更新されている．このような知識ベースをグラフ表現により変換し，マルコフ確立場によって推論を行うことでユーザの状態を推定する研究が存在する [8]．この方法では発話内で未観測である情報（単語・名詞等）を，外部知識ベースを用いて推論することで対話状態推定に利用できるようにしている．また，我々の以前の研究ではこうした知識グラフ上での推論を用いて，発話文のみから取得することが難しい対話状態推定に有効な情報を特徴量として得る方法を提案している [10]．具体的には，外部の知識として Wikidata を用い，ラベル伝搬法 [4] により推論することで，より有効な情報を含んだ特徴量ベクトルを生成している．

3 知識ベース上の推論による特徴ベクトルを用いたニューラルネットワークベースの対話状態推定

ユーザ発話では未観測の単語を推論により外部知識ベース上から見つけることで，より多くの意味的情報をもつ特徴量ベクトルを構築する．この特徴量ベクトルを識別モデルである対話状態推定器の入力とすることで，モデルの学習を行う．本節では，この特徴量ベクトルの構築方法と，対話状態推定器への適用について述べる．図 2 は提案法の概要図である．

3.1 知識ベース上の推論

知識ベース上の特徴量ベクトルの推論では，知識ベースから構築された知識グラフ上の観測ノードから，素性に加わるべき未観測ノードを予測する．ここで予測

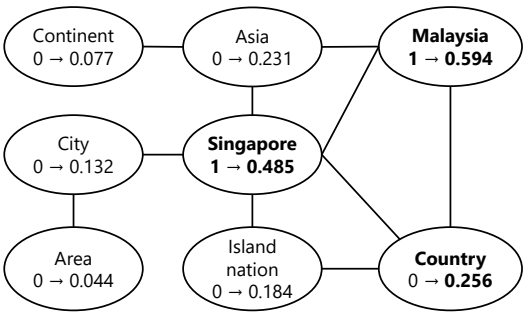

図 3: 推論した知識グラフの例.

されたノードの組み合わせを各発話の内容を表現したベクトルとする．

知識グラフは外部知識ベースである Wikidata を元に構築している．Wikidata から知識グラフへの変換では，データセットで観測した単語のうち Wikidata 内のエンティティとして存在するものをシードノードとする．また，シードノードと Wikidata 内で関係性のあるエンティティをグラフにおけるノードとして追加する．この際，1-hop 先のノードまで利用した．ここまでに得られた全てのエンティティ同士の関係を調べ，関係性が付与されていれば，それらのエンティティに相当するノードをエッジで繋げる．

このように構築した知識グラフ上でラベル伝搬法により発話文内で未観測のノードが発話から得られた情報に近いかを推論する．ラベル伝搬法は観測ノードのラベルから未観測ノードのラベルを推定するため，発話文では未観測であったノードのうち，グラフ上で近いものを観測されたノードと同じクラスとして推論することができる．未観測ノードのラベルは観測ノードに近いほど高い値を取り，遠いほど低い値を取る．図 3 は”Singapore” と”Malaysia” が観測語として与えられた時にラベル伝搬法を適用した結果であり，各ノードの左の値が入力値で右の値が予測値である．ラベル伝搬法では以下の目的関数を最小化する．

$$J(f) = \sum_{i=1}^{n}(y_i - f_i)^2 + \lambda \sum_{i<j} w_{i,j}(f_i + f_j)^2, \quad (1)$$

ここで y は入力値であり，f が予測値（出力値）である．w_{ij} はノード i とノード j のエッジの重みであり，λ は第一項と第二項のバランスをとる係数となる．予測したベクトル $\mathbf{f}$ を，

$$\mathbf{f} = \mathbf{y}(\mathbf{I} + \lambda \mathbf{L})^{-1} \quad (2)$$

のラプラシアン行列を用いた変形から求めることができる．また，ラベル伝搬法で対話履歴を考慮するため

に割引率として γ を定義した．この割引率を掛けた前発話の入力値と，現在の入力発話ベクトルとの和を取ることで，対話履歴を考慮した入力ベクトルとする．このように構築したベクトルの次元数が 51,548 次元となった．このままでは計算効率が低下するため，本研究では主成分分析（PCA）による次元圧縮をおこなった．この次元圧縮では累積寄与率を 1 に保つよう制約し，結果として対話状態推定器の入力は 2,500 次元に圧縮された．

3.2 知識グラフ特徴量ベクトルと全結合ニューラルネットワーク

提案する特徴量ベクトルは推論により多くの情報量を含む．ただし，この特徴量ベクトルは語順を考慮しておらず，既存の最も精度の高い推定器である CNN を用いた枠組みのように語順を考慮したモデルに適用できない．そのため，提案した特徴量ベクトルを入力とする対話状態推定器には全結合ニューラルネットワークを用いる．これは図 2 で示したように 3 層のニューラルネットから構成され，入力層と中間層の次元数は圧縮した特徴量ベクトルの次元数と同じになる．本稿ではこのモデルを IKG-FCNN と呼ぶ．

3.3 IKG-FCNN と CNN のアンサンブル

提案した特徴量ベクトルを用いた対話状態推定器はより多くの情報量を持つが，語順を考慮していない．この欠点を補うため，提案する IKG-FCNN と既存の CNN のアンサンブルによるモデルを構築した．アンサンブルの手法に応じて，1 つ目を *Ensemble-1*，もう一方を *Ensemble-2* とする．*Ensemble-1* では 2 つのモデルの線形補間をする．

$$\mathbf{y}_{ensemble_1} = \mathbf{y}_{fcnn} \times w_{fcnn} + \mathbf{y}_{cnn} \times w_{cnn} \quad (3)$$

式 (3) では $0 \leq w_{fcnn}, w_{cnn} \leq 1$ かつ $w_{fcnn} + w_{cnn} = 1$ である．*Ensemble-2* は各モデルの特徴量を用い，IKG-FCNN と CNN の中間層を直接結合した．具体的には IKG-FCNN の中間層（$\mathbf{h}_{fcnn}$）と CNN の max-pooling 後の特徴量ベクトル $\hat{c}_{cnn} + b$ を Rectified Linear Unit (ReLU) 活性化して結合した．

$$\mathbf{h}_{ensemble_2} = \mathbf{h}_{fcnn} \otimes ReLU(\hat{c}_{cnn} + b), \quad (4)$$

$\otimes$ はネットワークの連結である．式 (4) で得られた $\mathbf{h}_{ensemble_2}$ から出力層を全結合層で繋ぎ，シグモイド関数により出力 $\mathbf{y}_{ensemble_2}$ を得る．

$$\mathbf{y}_{ensemble_2} = \sigma(\mathbf{h}_{ensemble_2} * w + b). \quad (5)$$

全てのモデルは Adam optimizer で学習した．また，モデルの中間層では活性化関数として ReLU を使用し，出力層の誤差はシグモイド交差エントロピーを用いた．出力層では全ての Slot-Value の推定をおこなうため，出力層にシグモイド関数を用いた．また，バッチ正規化とドロップアウトを中間層とモデルの辺に適用した．この結果，最終的に各モデルでは出力層からシグモイドによる各 Slot-Value の確率が得られる．ただし，出力層では現在の発話のトピックに関連しない Slot-Value も推定するため，各トピックで必要な Slot-Value を参照するようにフィルタリングを加えた．

4 実験

DSTC4 のデータセットにより評価を行う．DSTC4 では *Accuracy* と *F-measure* の 2 つ評価指標が用いられている．*Accuracy* は副対話区間終了時点におけるフレームの完全一致率であり，Slot-Value の組が完全に一致していることを要求する．一方，*F-measure* は各 Slot-Value の組の一致を *precision* と *recall* の調和平均で計算したものである．

本実験では 4 つのモデル（IKG-FCNN，CNN，*Ensemble-1*，*Ensemble-2*）による対話状態推定の精度の比較をおこなう．これらのハイパーパラメータは開発セットを用いて以下の通りに決定した．FCNN は Adam の α が 0.000025，バッチサイズが 40，出力層のシグモイドにおける閾値が 0.2 とした．CNN では Adam の α が 0.00005，重み減衰率が 0.000001，フィルタ幅が 1，出力チャンネル数が 1500，バッチサイズが 50，出力層の閾値が 0.2 とした．*Ensemble-1* では各モデルの重みが 0.5，出力層の閾値が 0.2 とした．*Ensemble-2* では Adam の α が 0.00005，重み減衰率が 0.000001，フィルタ幅が 1，出力チャンネル数が 500，バッチサイズが 50，出力層の閾値が 0.3 とした．くわえて DSTC4 で機械学習による最も高い精度を達成した対話状態推定器 (MSIIP) を比較のため併記する [7]．表 1 が各モデルの *Accuracy*，*precision*，*recall*，*F-measure* の結果を示している．

Ensemble-1 の *F-measure* が IKG-FCNN より 2.8%，MSIIP より 1.6%ほど高く，最高値を実現した．このモデルは *precision*，*recall* 共に IKG-FCNN と CNN を上回っており，それぞれの良い点を上手く学習することができていると考えられる．一方で，*Ensemble-2* の *F-measure* は IKG-FCNN から上がっておらず，CNN の結果を受けて IKG-FCNN より *precision* が高くなったのに対して，*recall* が下がってしまったと考えられる．

表 2 は，Transcription の発話文に対して各モデルが実際に推定した対話状態の例である．この例では各アン

	IKG-FFNN	CNN	Ensemble-1	Ensemble-2	MIISP
Accuracy	0.0445	0.0322	0.0578	0.0540	**0.0697**
Precision	0.3766	0.4339	**0.5167**	0.4690	0.4634
Recall	0.3102	0.2816	0.3307	0.3086	**0.3335**
F-measure	0.3758	0.3415	**0.4033**	0.3723	0.3878

表 1: 実験結果の比較表.

Transcription	Gold Standard	IKG-FCNN	CNN	Ensemble-1	Ensemble-2
also for certain rides in the Universal Studio, there's a height limit.	**PLACE**: 'Universal Studios Singapore' **ACTIVITY**: 'Amusement ride' **INFO**: 'Restriction'	**PLACE**: 'Universal Studios Singapore'	**PLACE**: 'Universal Studios Singapore'	**PLACE**: 'Universal Studios Singapore' **ACTIVITY**: 'Amusement ride' **INFO**: 'Restriction'	**PLACE**: 'Universal Studios Singapore' **ACTIVITY**: 'Amusement ride' **INFO**: 'Restriction'

表 2: Utterance-id 566 in Dialog #21 の対話状態フレーム.

サンブルモデルがいずれも正解のフレームと完全一致しており，元となった IKG-FCNN や CNN 単体では推定できていない Slot-Value が存在した (**ACTIVITY**, **INFO**). この結果からもアンサンブルモデルでは各モデルを合わせることで相乗効果があったと推定できる.

5 まとめ

本研究では知識グラフ上の推論により発話文そのものからは抽出が難しい特徴量の抽出を行い，この特徴量ベクトルを用いてニューラルネットワークによる対話状態推定を行った．また，この推定器と CNN ベースの推定器をアンサンブルする方法も提案した．その結果，線形補間によるアンサンブルモデルが，その他の提案したモデルや DSTC4 における機械学習による最高精度の推定器の結果を上回った．ただし，提案した IKG-FCNN では元の発話内の語順の系列情報を考慮していないため，このような系列情報を扱えるようにすることが今後の課題である.

6 謝辞

本研究開発の一部は総務省 SCOPE(受付番号 152307004) の委託を受けたものです.

参考文献

[1] Sören Auer, Cristian Bizer, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735, 2007.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of ACS SIGMOD International Conference on Management of Data*, pages 1247–1250, 2008.

[3] Matthew Henderson. Machine learning for dialog state tracking: A review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language*, 2015.

[4] Tsuyoshi Kato, Hisashi Kashima, and Masashi Sugiyama. Robust label propagation on multiple networks. In *IEEE Transactions on Neural Networks*, volume 20, pages 35–44, 2009.

[5] Seokhwan Kim, Luis D'Haro, Rafael Banchs, Matthew Henderson, Jason Williams, and Koichiro Yoshino. *Dialog State Tracking Challenge 4*. http://www.colips.org/workshop/dstc4/, 2015.

[6] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of Conference on Empirical Methods on Natural Language Processing*, 2014.

[7] Miao Li and Ji Wu. The msiip system for dialog state tracking challenge 4. In *Dialogues with Social Robots*, pages 465–474. Springer Singapore, 2017.

[8] Yi Ma, Paul Crook, Rushi Sarikayu, and Eric Fosler-Lussier. Knoewledge graph inference for spoken dialog system. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 5346–5305, 2015.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed represenatational of words and phrases and their compositionality. In *Proceedings of Advances in Nueral Information Processing System*, pages 3111–3119, 2013.

[10] Yukitoshi Murase, Koichiro Yoshino, Masahiro Mizukami, and Satoshi Nakamura. Feature inference based on label propagation on wikidata graph for dialogue state tracking. In *International Workshop on Spoken Dialogue System Technology*, pages 1–12, 2017.

[11] Hongjie Shi, Takashi Ushino, Mitsuruand Endo, Katsuyoshi Yamagami, and Noriaki Horii. A multichannel convolutional neural network for cross-language dialog state tracking. In *Proceedings of the Spoken Language Technology 2016*, pages 559–564, 2016.

[12] Deny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. In *Communications of the ACS*, pages 78–85, 2014.

[13] Jason Williams, Antonie Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *Proceedings of the Special Interest Group on Discource and Dialogue 2013 Conference*, pages 404–413, 2013.

[14] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, 2010.

[15] Lukas Zilka and Filip Jurcicek. Incremental lstm-based dialog state tracker. In *Automatic Speech Recognition and Understanding*, pages 757–762, 2015.