

EMNLP 2016 参加報告

2016/12/21

若林 啓 (筑波大学)

Agenda

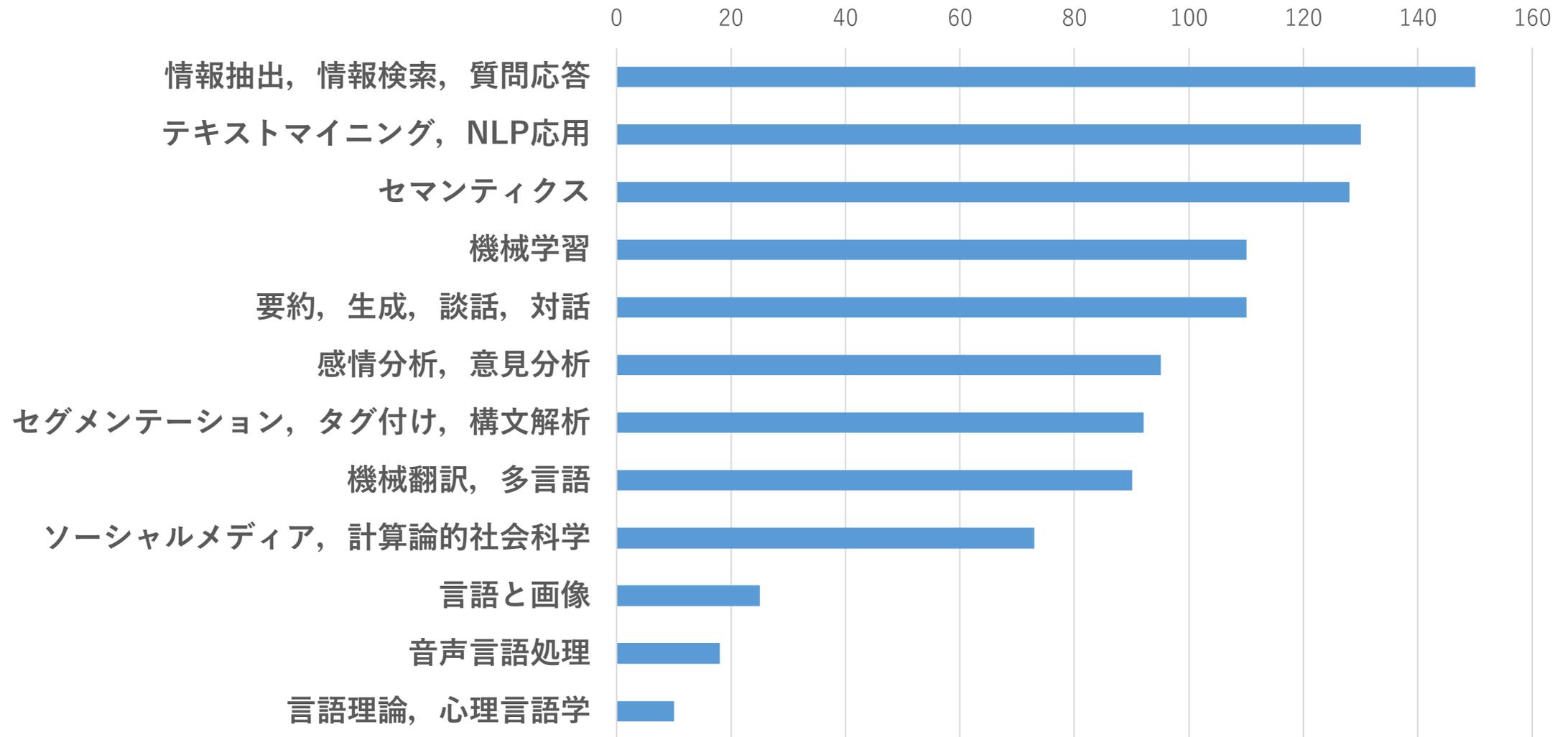
- EMNLP 2016 とは
- 3件の招待講演の内容
 - 語用論の話
 - ロボットの自然言語インタフェースの話
 - Addressee detectionの話
- 対話システムに関する研究発表の紹介
 - 発話選択型手法について
 - 発話生成型手法について
 - 対話システムの自動評価方法について

EMNLPとは

- Conference on Empirical Methods in Natural Language Processing
 - EMNLP 2016 はAustin, Texasで開催（米国での開催は4回目）
 - 自然言語処理系の著名な国際会議のひとつ
 - ACL special interest group on linguistic data and corpus-based approaches



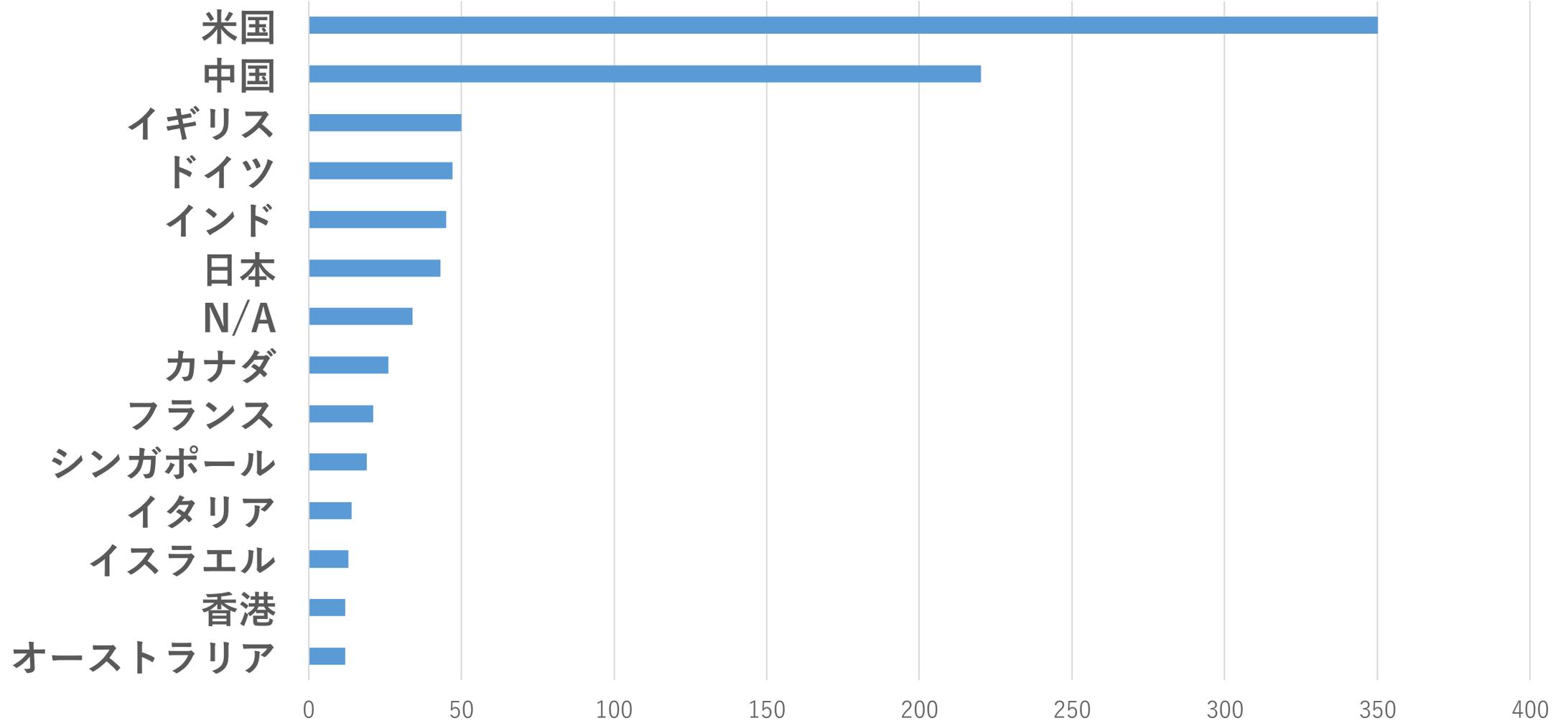
分野別の投稿数 (概数)



EMNLP 2016 採択率

| | Long Papers | Short Papers | 合計 |
|-------------|---------------|---------------|---------------|
| 投稿数 | 747 | 438 | 1,185 |
| 査読された投稿数 | 687 | 400 | 1,087 |
| 採択数 | 177 | 87 | 264 |
| 採択率 | 25.76% | 21.75% | 24.29% |
| TACL Papers | | | 9 |
| 発表論文数 | | | 273 |

国別の投稿数（概数）



会議全体の構成

- 本会議は 11/2, 3, 4 の 3 日 日
 - 11/1 にチュートリアル+ワークショップ
 - 11/5 にもワークショップ
- 招待講演3件
- 24の口頭発表セッション
- [Half minute発表 + ポスター発表]が2回
 - HMMは今年初めての試み
 - ロングペーパーのポスター発表者が登壇
- ベストペーパーセッション



Agenda

- EMNLP 2016 とは
- **3件の招待講演の内容**
 - 語用論の話
 - ロボットの自然言語インタフェースの話
 - Addressee detectionの話
- 対話システムに関する研究発表の紹介
 - 発話選択型手法について
 - 発話生成型手法について
 - 対話システムの自動評価方法について

招待講演1



• Learning in extended and approximate Rational Speech Acts models

- Christopher Potts (Stanford University)
- 主に語用論 (pragmatics) の話

• Rational Speech Actモデル

- 相手がどう解釈するかを考慮して使う語彙を決める
- TUNAデータセットの話
 - 深層学習の適用
- 写真ではなく色の場合
- 協調ゲームにおけるコミュニケーション

Pragmatic Speaker

Pragmatic Listener

| | <i>beard</i> | <i>glasses</i> | <i>tie</i> |
|---|--------------|----------------|------------|
|  | .67 | .33 | 0 |
|  | 0 | .6 | .4 |
|  | 0 | 0 | 1 |

| <i>beard</i> |  |  |  |
|----------------|--|--|--|
| <i>beard</i> | 1 | 0 | 0 |
| <i>glasses</i> | .5 | .5 | 0 |
| <i>tie</i> | 0 | .33 | .67 |

招待講演2



• Learning Models of Language, Action and Perception for Human-Robot Collaboration

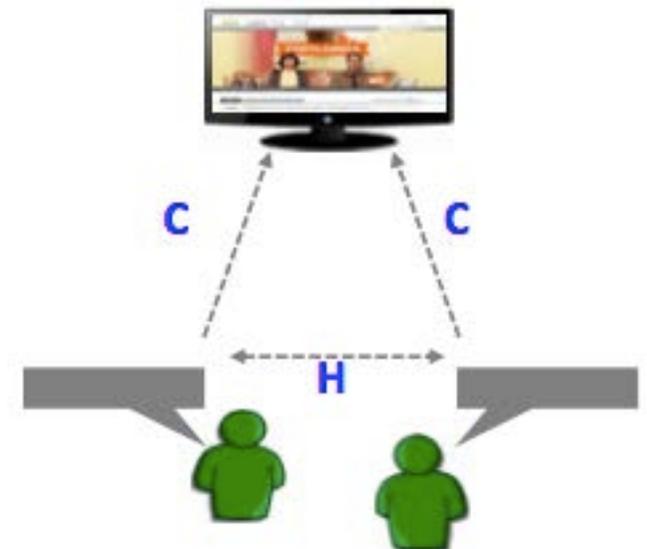
- Stefanie Tellex (Brown University)
- ロボットを自然言語で操作する話
 - **Partially Observable Markov Decision Process (POMDP)**
 - ユーザ発話の意図理解
 - 何をしてほしいのか
 - どのタイヤ（トラック）？
 - ジェスチャや状況の理解
 - ロボットが取るべきアクションとプランニング
 - 何を取りに行くか、どう移動するか
 - 指示が曖昧なら**ユーザに質問**する
 - 統一的に扱えるモデルとして有望



招待講演3



- **You Talking to Me? Speech-based and multimodal approaches for human versus computer addressee detection**
 - Andreas Stolcke (Microsoft Research)
- 合図 (“Hey Siri” など) 無しでユーザに反応する手法の話
 - **Addressee detection task**
 - 言語特徴量よりも **音響特徴量が有効**
 - システムに向けて話すとき，人はより明瞭に話す
 - Monaco Dataset (非公開)
 - 2-3人で音声対話エージェントとクイズゲーム
 - マイク音声，カメラ画像を記録
 - 画像情報はあまり役に立たない
 - 相関はあるが音響特徴量に比べて弱い



招待講演から感じた今後のトレンド(?)

- 自然言語によるコミュニケーション
 - これまで困難だった様々な課題への挑戦
 - 文脈・状況の考慮, 意味のグラウンディング, 韻律がもつニュアンスの認識, マルチモーダル, 質問応答, 知識・常識の利用, …
 - 深層学習をはじめとして基盤技術が高度化してきた
- システム側からの自然言語による情報伝達
 - 対話インタフェースの応用範囲が広がってきている
 - デバイスの発達, ロボットの普及などのインフラ面の充実

Agenda

- EMNLP 2016 とは
- 3件の招待講演の内容
 - 語用論の話
 - ロボットの自然言語インタフェースの話
 - Addressee detectionの話
- **対話システムに関する研究発表の紹介**
 - 発話選択型手法について
 - 発話生成型手法について
 - 対話システムの自動評価方法について

対話システムに関する研究発表

発話
選択型

- **Multi-view Response Selection for Human-Computer Conversation**
- **Addressee and Response Selection for Multi-Party Conversation**

発話
生成型

- **Deep Reinforcement Learning for Dialogue Generation**
- **Conditional Generation and Snapshot Learning in Neural Dialogue Systems**

サブ
モジュール

- **Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems**
- **Nonparametric Bayesian Models for Spoken Language Understanding**

評価方法

- **How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation**

コーパス
構築

- **The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues**

対話システムに関する研究発表

発話
選択型

- **Multi-view Response Selection for Human-Computer Conversation**
- **Addressee and Response Selection for Multi-Party Conversation**

発話
生成型

- **Deep Reinforcement Learning for Dialogue Generation**
- **Conditional Generation and Snapshot Learning in Neural Dialogue Systems**

サブ
モジュール

- **Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems**
- **Nonparametric Bayesian Models for Spoken Language Understanding**

評価方法

- **How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation**

コーパス
構築

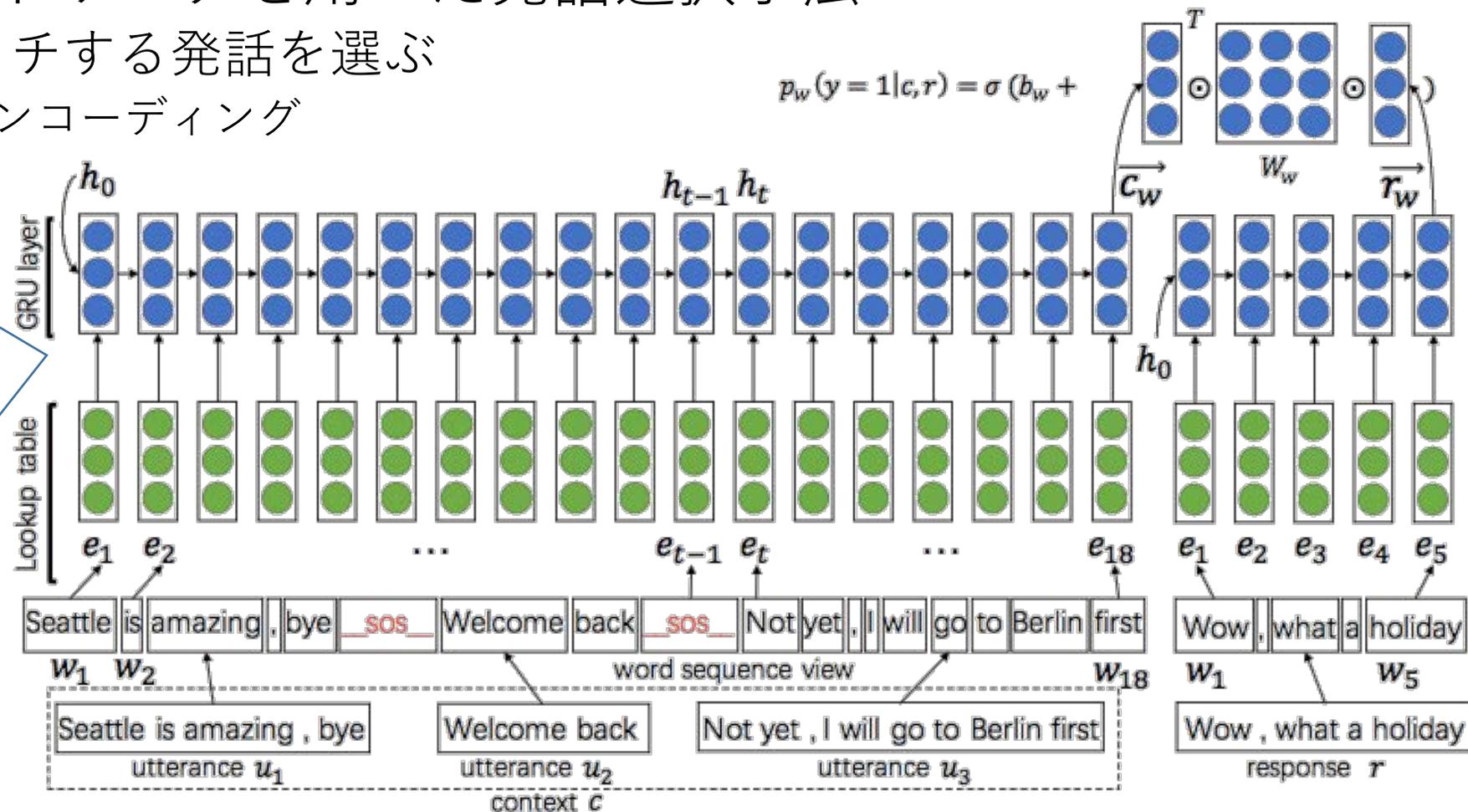
- **The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues**

Multi-view Response Selection for Human-Computer Conversation

- ニューラルネットワークを用いた発話選択手法
 - 最も文脈とマッチする発話を選ぶ
 - RNNによるエンコーディング

直近の発話の影響が強くなりすぎる傾向

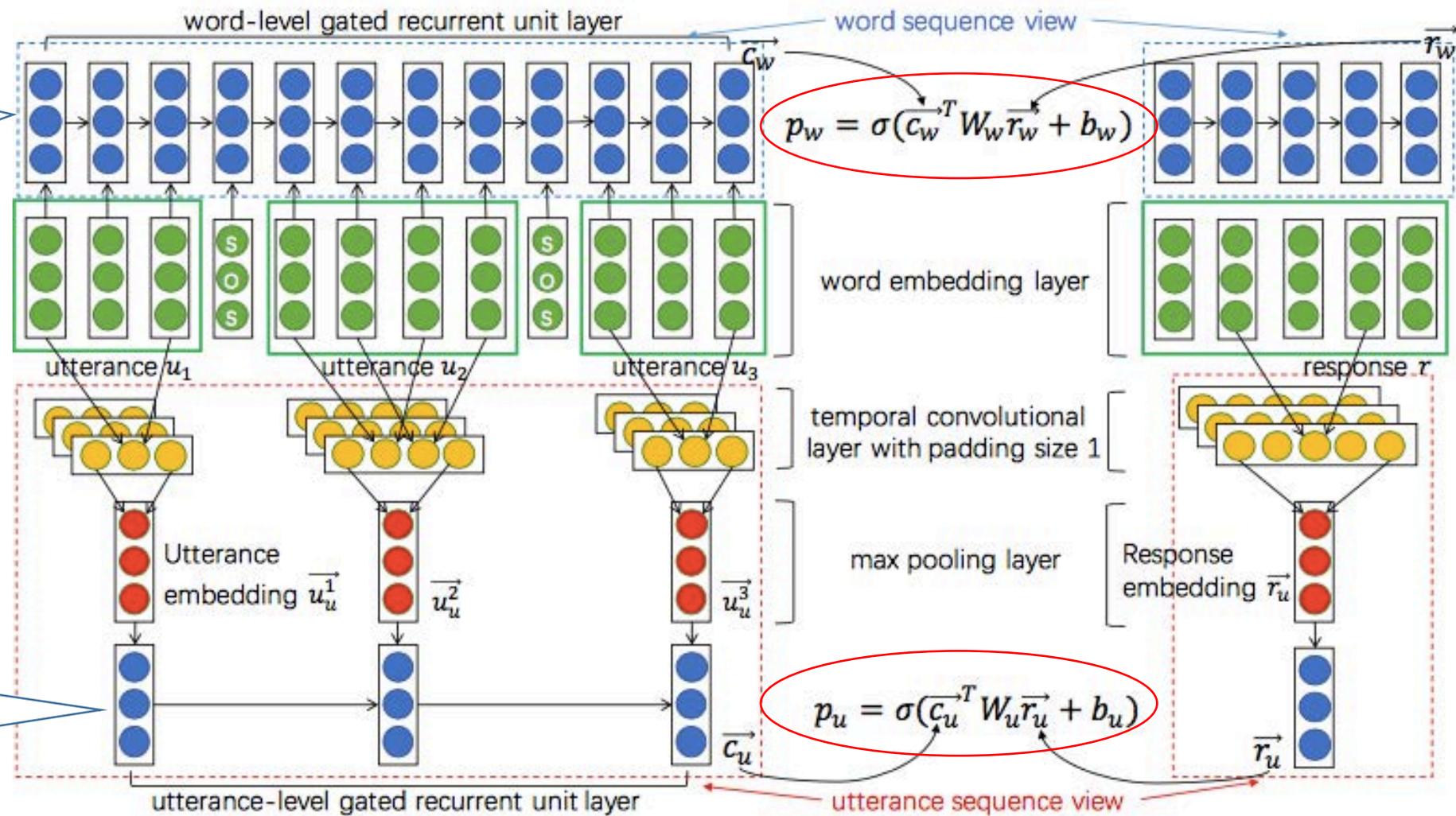
「1発話1ベクトル」のRNNを別途学習する



Utterance Embedding

従来通りの
単語レベルRNN

新たに加えた
発話レベルRNN



- 損失関数の設定

- \mathcal{L}_D : 単語RNNと発話RNNの不一致が少ない方がよい
- \mathcal{L}_L : 各RNNがそれぞれ正解のラベルを当てる確率
 - 加えて, 正規化項

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_L + \frac{\lambda}{2} \|\theta\|$$

$$\mathcal{L}_D = \sum_i (p_w(l_i) \bar{p}_u(l_i) + p_u(l_i) \bar{p}_w(l_i))$$

$$\mathcal{L}_L = \sum_i (1 - p_w(l_i)) + \sum_i (1 - p_u(l_i))$$

- Ubuntuコーパスでの評価, 正解が上位k件に含まれる割合が向上

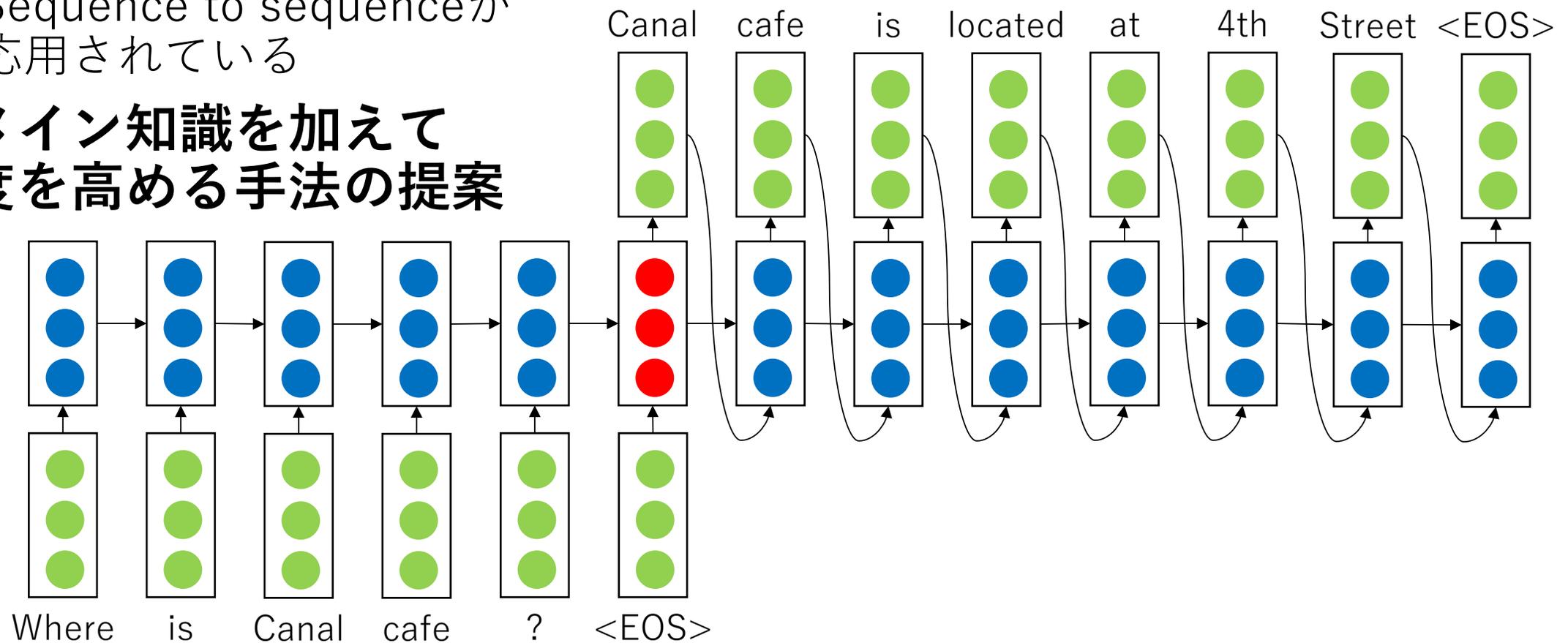
| Model/Metrics | 1 in 10 R@1 | 1 in 10 R@2 | 1 in 10 R@5 | 1 in 2 R@1 |
|-----------------------------------|---------------|---------------|---------------|---------------|
| Random-guess | 10% | 20% | 50% | 50% |
| TF-IDF | 41.0% | 54.5% | 70.8% | 65.9% |
| Word-seq-LSTM (Lowe et al., 2015) | 60.40% | 74.50% | 92.60% | 87.80% |
| Word-seq-GRU | 60.85% | 75.71% | 93.13% | 88.55% |
| Utter-seq-GRU | 62.19% | 76.56% | 93.42% | 88.83% |
| Multi-view | 66.15% | 80.12% | 95.09% | 90.80% |

Conditional Generation and Snapshot Learning in Neural Dialogue Systems

- ニューラルネットワークを用いた発話生成手法

- Sequence to sequenceが応用されている

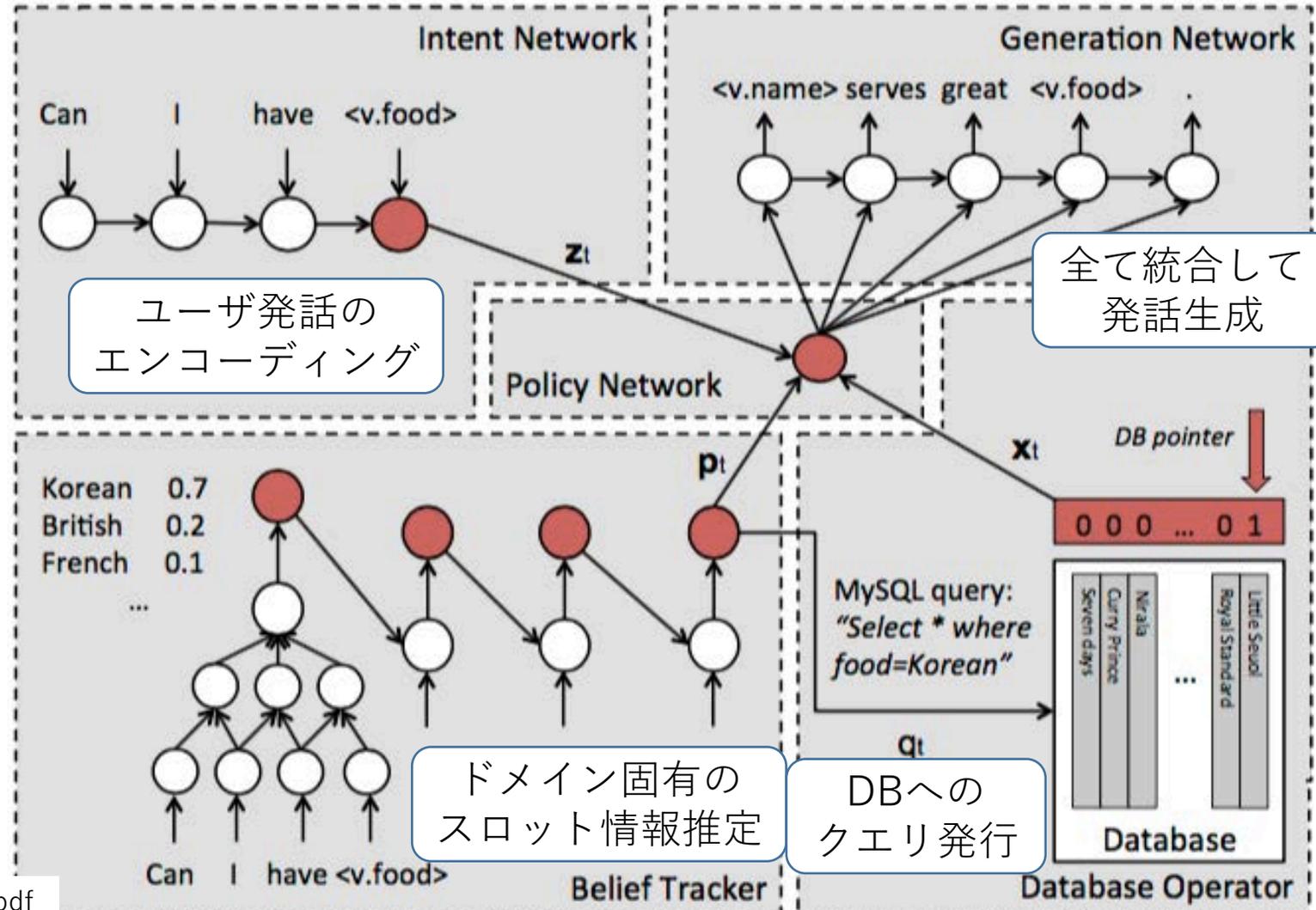
- ドメイン知識を加えて精度を高める手法の提案**



知識ベースへの問い合わせを含むモデル

料理屋検索を想定

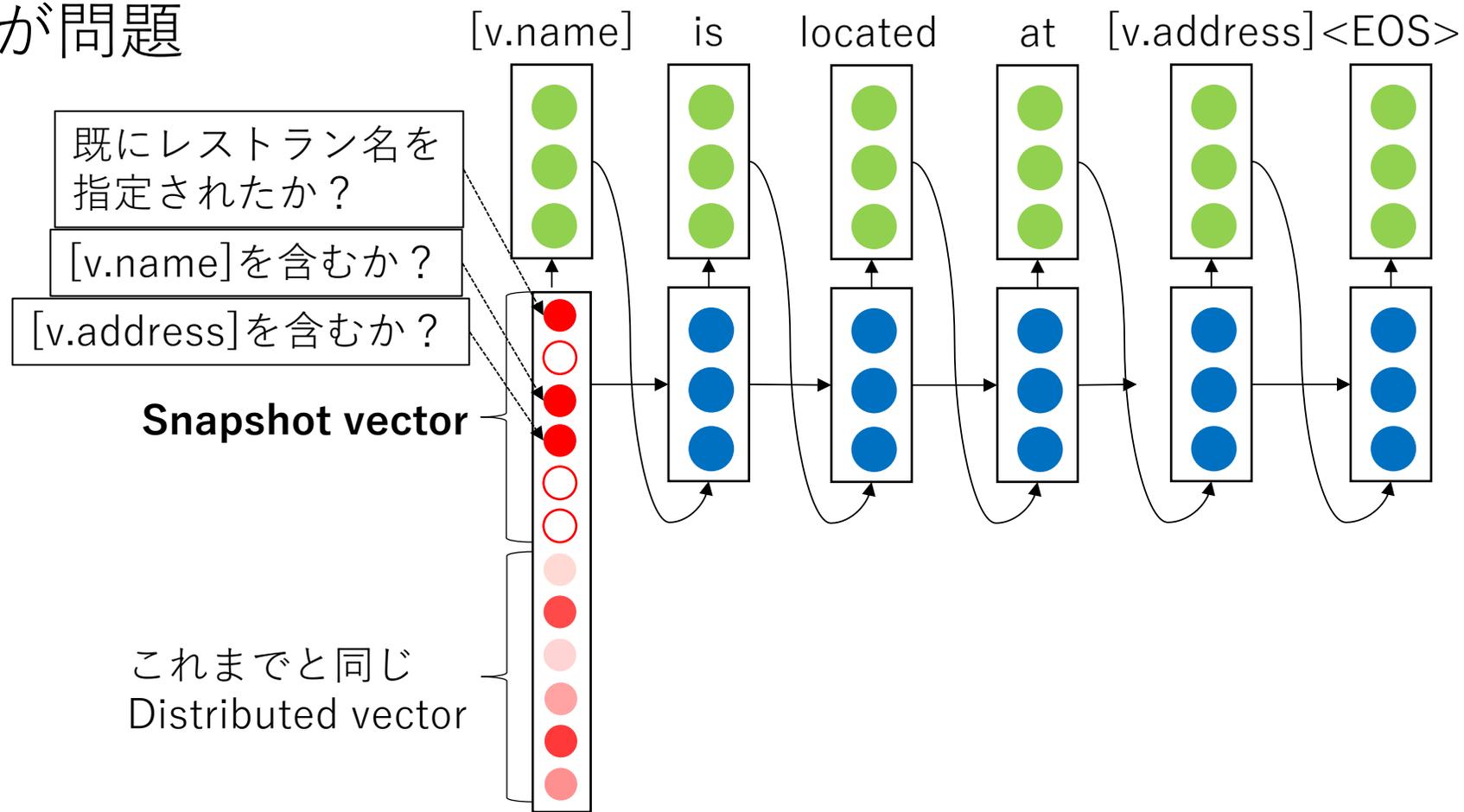
- スロット情報の認識
 - 何料理が食べたいか
 - 場所はどこか
 - 価格帯はどれくらいか
- DBには必要な情報がある
 - スロット値が分かれば料理屋を検索できる
- 発話生成は2段階
 - RNNによる生成ではプレースホルダを出力
 - そこにDBの検索結果を代入して発話が完成



Snapshot Learning

- 中間表現をなかなかうまく学習してくれないのが問題

- 中間表現の教師情報
 - 一部の次元の意味を予め決めてしまう
- これにより精度が改善できる



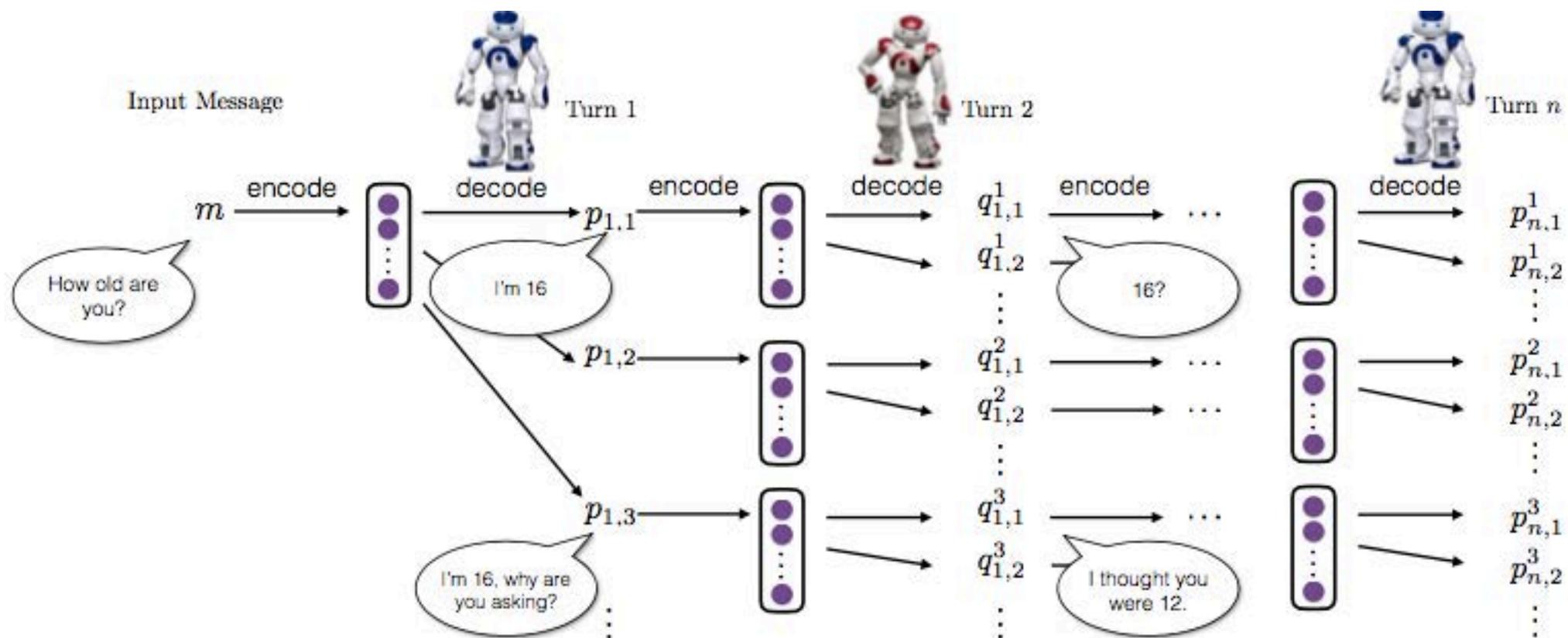
Deep Reinforcement Learning for Dialogue Generation

- 発話生成型の問題の1つは「つまらない発話が生成されやすい」
 - エージェント同士を会話させてみた例が左下図
 - すぐに「会話のブラックホール」に落ちることが分かる
- **強化学習の枠組みを適用する**
 - **会話が続きやすい発話生成を行う手法の提案（右下図）**

| Baseline mutual information model (Li et al. 2015) | Proposed reinforcement learning model |
|--|---|
| A: Where are you going? (1) | A: Where are you going? (1) |
| B: I'm going to the restroom. (2) | B: I'm going to the police station. (2) |
| A: See you later. (3) | A: I'll come with you. (3) |
| B: See you later. (4) | B: No, no, no, no, you're not going anywhere. (4) |
| A: See you later. (5) | A: Why? (5) |
| B: See you later. (6) | B: I need you to stay here. (6) |
| ... | A: I don't know what you are talking about. (7) |
| | ... |

Deep Reinforcement Learning

- 状態空間をRNNのベクトルで表現したモデル
 - 将来の数ステップにおける報酬の期待値を最大化するように学習



報酬関数の設定 (1. 2. 3. の重み和とする)

1. 頻繁に見られる「面白くない返答」を人手で設定
 - “I don't know what you are talking about”, “I have no idea”, など8種類
 - これらを使ったらマイナスの報酬

2. 新しい情報を含むような発話に報酬を与える

$$r_2 = -\log \cos(h_{p_i}, h_{p_{i+1}}) = -\log \frac{h_{p_i} \cdot h_{p_{i+1}}}{\|h_{p_i}\| \|h_{p_{i+1}}\|}$$

- h_{p_i} は発話 p_i を RNN でエンコードしたベクトル
- 直前の発話とのコサイン類似度が低いほどよい

3. 文章としてまともな発話かどうか

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$$

- RNN からの生成確率が高いほどよい

- エージェント同士の会話がより長く続くようになった
 - 比較手法の mutual information はSeq2seqの改良

| Model | # of simulated turns |
|--------------------|----------------------|
| SEQ2SEQ | 2.68 |
| mutual information | 3.40 |
| RL | 4.48 |

- 従来手法よりも会話が続きやすそうな発話が生成できている

| Input Message | Mutual Information Model | RL |
|----------------------------------|--|---|
| How old are you? | I'm 16. | I'm 16, why are you asking? |
| What's your full name? | I have no idea. | What's yours? |
| I don't want to go home tonight. | Really? | Why? |
| Do you have any feelings for me? | I don't know what you are talking about. | I don't want to hurt your feelings. |
| How much time do you have here? | Not long enough. Sorry, sir. | Ten seconds. |
| Shall we get started? | Of course! | Yes. We've got a lot of work to do here. |
| Do you play football? | No, i don't | Yes. I love football! |
| We'd have to talk to him. | I mean, he's a good guy | About what ? |
| How come you never say it? | Because I don't want to hurt you. | I don't think it's a good idea to say it. |

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

- 対話システムの自動評価手法に関する問題
 - あらかじめ用意されたGround-Truth応答との類似度に基づいている
 - 翻訳の評価に使われるBLEUや、要約の評価に使われるROUGEなど

- 右図のようなModel ResponseのBLEUスコアは0になる

Context of Conversation

Speaker A: Hey John, what do you want to do tonight?

Speaker B: Why don't we go see a movie?

Ground-Truth Response

Nah, I hate that stuff, let's do something active.

Model Response

Oh sure! Heard the film about Turing is out!

- 人手評価との乖離を明らかにする

評価される評価手法

- 単語一致ベース手法
 - BLEU-N : N-gramの一致率
 - METEOR : 明示的に単語をアライメントした上での単語一致のF-measure
 - ROUGE-L : Longest Common Subsequence (LCS) に基づくF-measure
- Embeddingベース手法
 - Greedy Matching : 単語のword embeddingのコサイン類似度の和
 - 各単語は, 最もコサイン類似度の高い単語を相手に選ぶ
 - Embedding Average : 文章のベクトルは構成単語のベクトルの平均とする
 - そのコサイン類似度を取る
 - Vector Extrema : 文章のベクトルの各成分を, 構成単語の最大値とする
 - そのコサイン類似度を取る
- いずれも, Ground-Truthとどれだけ似ているかを測る

実験に用いた応答手法

- 発話選択型手法
 - TF-IDF：ベクトルの類似度が高い発話を選択
 - Dual Encoder：文脈をRNNにかけて得られるEmbeddingの類似度が高い発話を選択
- 発話生成手法
 - LSTM：最も確率の大きい状態にグリーディに遷移して単語予測し続ける
 - Hierarchical Recurrent Encoder-Decoder (HRED)：1 発話 1 ベクトルに対応したRNNによるEncoder-Decoder

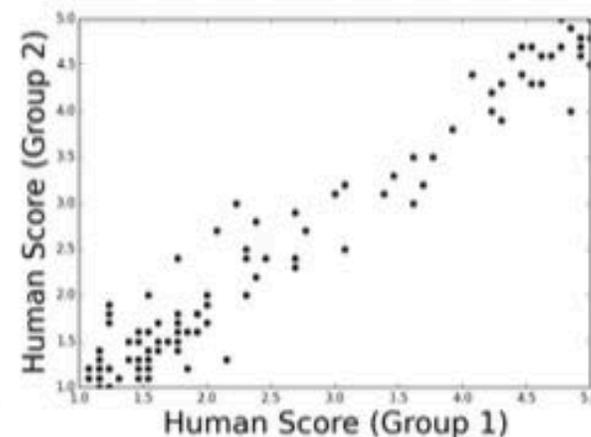
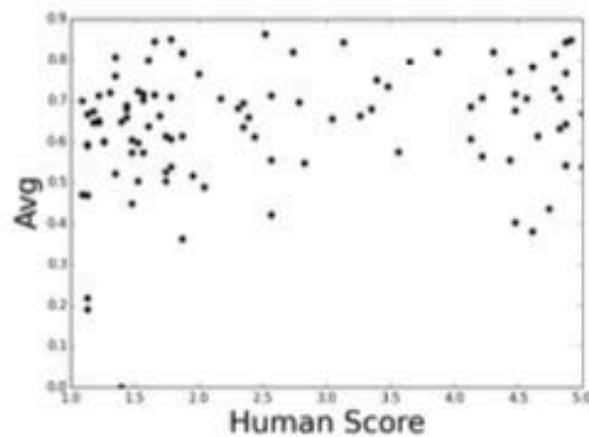
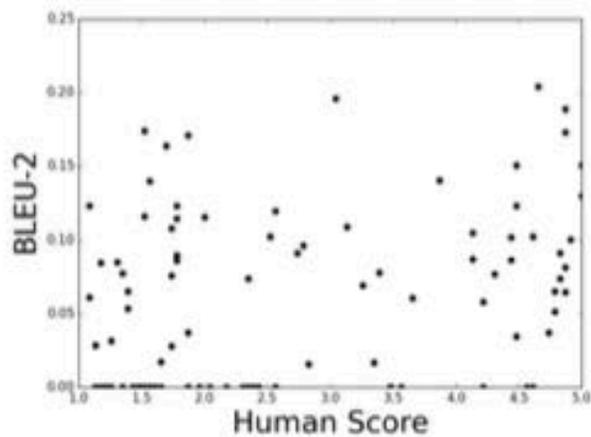
| | Ubuntu Dialogue Corpus | | | Twitter Corpus | | |
|---------|------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Embedding Averaging | Greedy Matching | Vector Extrema | Embedding Averaging | Greedy Matching | Vector Extrema |
| R-TFIDF | 0.536 ± 0.003 | 0.370 ± 0.002 | 0.342 ± 0.002 | 0.483 ± 0.002 | 0.356 ± 0.001 | 0.340 ± 0.001 |
| C-TFIDF | 0.571 ± 0.003 | 0.373 ± 0.002 | 0.353 ± 0.002 | 0.531 ± 0.002 | 0.362 ± 0.001 | 0.353 ± 0.001 |
| DE | 0.650 ± 0.003 | 0.413 ± 0.002 | 0.376 ± 0.001 | 0.597 ± 0.002 | 0.384 ± 0.001 | 0.365 ± 0.001 |
| LSTM | 0.130 ± 0.003 | 0.097 ± 0.003 | 0.089 ± 0.002 | 0.593 ± 0.002 | 0.439 ± 0.002 | 0.420 ± 0.002 |
| HRED | 0.580 ± 0.003 | 0.418 ± 0.003 | 0.384 ± 0.002 | 0.599 ± 0.002 | 0.439 ± 0.002 | 0.422 ± 0.002 |

人手評価との順位相関係数

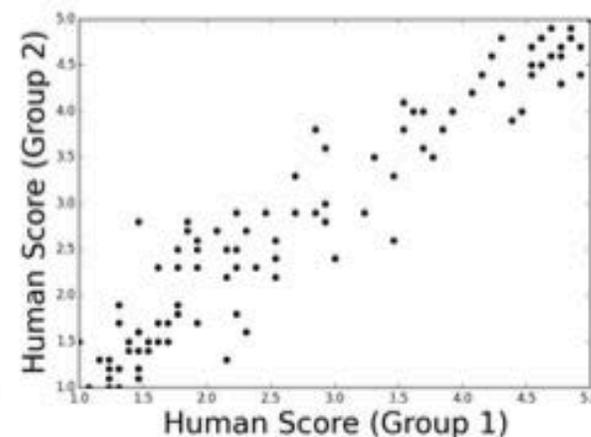
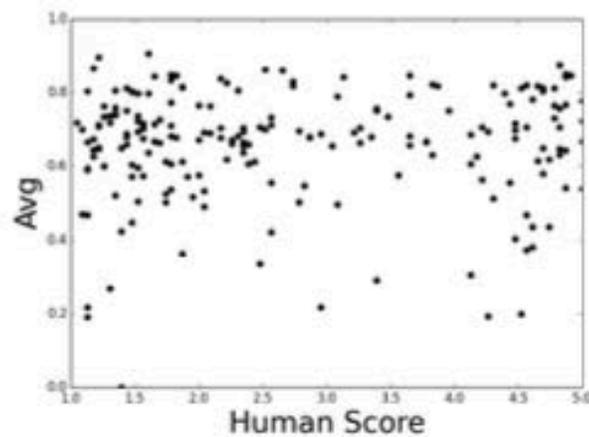
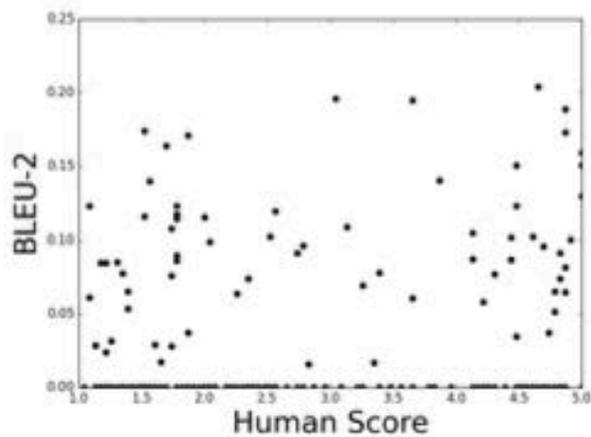
- 25人の評価者が1から5のスケールで100発話を評価
 - **自動評価手法との順位相関は非常に弱い**ことが分かった

| Metric | Twitter | | | | Ubuntu | | | |
|---------|----------|---------|---------|---------|----------|---------|-----------|---------|
| | Spearman | p-value | Pearson | p-value | Spearman | p-value | Pearson | p-value |
| Greedy | 0.2119 | 0.034 | 0.1994 | 0.047 | 0.05276 | 0.6 | 0.02049 | 0.84 |
| Average | 0.2259 | 0.024 | 0.1971 | 0.049 | -0.1387 | 0.17 | -0.1631 | 0.10 |
| Extrema | 0.2103 | 0.036 | 0.1842 | 0.067 | 0.09243 | 0.36 | -0.002903 | 0.98 |
| METEOR | 0.1887 | 0.06 | 0.1927 | 0.055 | 0.06314 | 0.53 | 0.1419 | 0.16 |
| BLEU-1 | 0.1665 | 0.098 | 0.1288 | 0.2 | -0.02552 | 0.8 | 0.01929 | 0.85 |
| BLEU-2 | 0.3576 | < 0.01 | 0.3874 | < 0.01 | 0.03819 | 0.71 | 0.0586 | 0.56 |
| BLEU-3 | 0.3423 | < 0.01 | 0.1443 | 0.15 | 0.0878 | 0.38 | 0.1116 | 0.27 |
| BLEU-4 | 0.3417 | < 0.01 | 0.1392 | 0.17 | 0.1218 | 0.23 | 0.1132 | 0.26 |
| ROUGE | 0.1235 | 0.22 | 0.09714 | 0.34 | 0.05405 | 0.5933 | 0.06401 | 0.53 |
| Human | 0.9476 | < 0.01 | 1.0 | 0.0 | 0.9550 | < 0.01 | 1.0 | 0.0 |

100発話の散布図



(a) Twitter



(b) Ubuntu

よりよい自動評価手法の確立に向けて

- 今回の実験はオープンドメインだった
 - 制約の強いタスクの場合, BLEUとの相関は強まるだろう
- 正解が大量に用意できればよいのではないか
 - 正解が多数あるときの評価指標の提案がFuture workとのこと
- Embedding-based metricsが今後発展の見込みあり
 - よりよいモデルを使えば, よりよい評価指標になるはず
 - 会話の文脈を評価モデルに加えることも検討できる