単語の分散表現および音素列の類似性を考慮した単語アラインメントに基づく教師なし Entity Linking

邊土名 朝飛　友松 祐太　杉山 雅和　戸田 隆道　東 佑樹　下山 翔
株式会社 AI Shift
{hentona_asahi, tomomatsu_yuta, sugiyama_masakazu,
toda_takamichi, azuma_yuki, sho_shimoyama_xb}@cyberagent.co.jp

概要

本研究では，多様なパターンのユーザ発話やエントリを考慮した，音声認識誤りに頑健な教師なし Entity Linking 手法を提案する．提案手法は，単語の分散表現と音素列を利用してアラインメントをとることで，意味的類似性と音韻的類似性の両方を考慮する．手法の妥当性を検証するため，自社で運用している音声対話システムのデータを利用し，テキスト間類似度を測る複数の手法との間で比較実験を行った．実験の結果，提案手法は従来手法よりも高い性能を示し，多様なパターンのユーザ発話やエントリを考慮できることがわかった．

1　はじめに

Entity Linking とは，テキスト中に含まれる Entity を認識し，知識ベース上のエントリと紐付けるタスクである．タスク指向型の音声対話システムにおいては，前段で自動音声認識（Automatic Speech Recognition; ASR）システムを用いてユーザ発話をテキスト化し，後段の言語理解（Natural Language Understanding; NLU）モジュールで Entity Linking を行う構成がとられていることが多い．

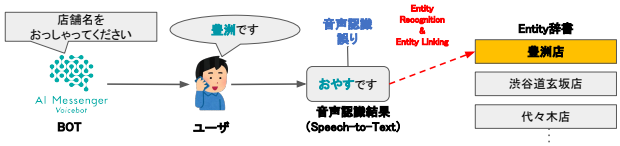

図 1　音声対話システムにおける Entity Linking

しかし，近年の汎用 ASR システムの認識性能は非常に高い性能を有しているものの，依然としてドメイン固有の単語は誤認識しやすいという問題がある．ドメインごとに音声認識モデルを構築することで認識誤りを低減することは可能であるが [1]，ドメインごとに十分な量の学習データを用意してモデルを学習するためには多大なコストがかかる．この問題に加えて，環境音などのノイズや音質劣化，イントネーションの差異などの要因により認識性能はさらに低下する．また，ユーザ発話やエントリには多様な表記方式や言い間違いのパターンが存在するため，より Entity Linking を難しいものにしている（表 1 参照）．

表 1　ユーザ発話，エントリの例

種別	例
省略	エントリ：サイバーエージェント 発話　　：サイバー
単語誤り	エントリ：〇〇公園 発話　　：〇〇広場
転置	エントリ：〇〇 渋谷店 発話　　：渋谷〇〇店
関連フレーズ	エントリ：Abema Towers 発話　　：サイバーエージェント
表記違い	Abema Towers，アベマタワーズ

本研究では，多様なパターンのユーザ発話やエントリを考慮した，音声認識誤りに頑健な教師なし Entity Linking 手法を提案する．提案手法は，単語分散表現で意味的類似度を，単語音素列で音韻的類似度を考慮してアラインメントをとり，ユーザ発話と各エントリ間の類似度を計算することで Entity Linking を実現する[1)]．実験では，自社で運用している音声対話システムのログデータを使用し，アブレーション実験およびテキスト間類似度を測る複数の教師なし手法との間で比較実験を行うことで提案手法の有効性を示す．

1)　本来であれば Entity Linking の前にテキスト中から Entity を抽出する必要があるが，実験に用いた対話データではほとんどのユーザが Entity のみ発話していたため，直接 Entity Linking を行っている．

2 関連研究

Raghuvanshi らは，ASR の認識結果をデータベース上の人名と紐付けることを目的とした教師なし手法を提案した [2]．この手法は，テキストと音素の類似度を利用することで音声認識誤りに頑健な人名検索を行うことができる．本研究では，Raghuvanshi らと同様に音素情報を用いるとともに，単語分散表現も利用することで意味的類似度も考慮した Entity Linking を行う．

単語ベクトルを利用して 2 つのテキスト間の意味的類似性を計算する主流のアプローチとして，最適輸送に基づいたアラインメント手法がある．代表的な手法としては Word Mover's Distance (WMD)[3] が挙げられ，その他にも超球面上で最適輸送を行う Word Rotator's Distance (WRD)[4] や，不均衡最適輸送 (Unbalanced Optimal Transport; UOT) を導入した Lazy Earth Mover's Distance (Lazy-EMD)[5] などが提案されている．本研究では，ユーザ発話とエントリ間の類似度を計算するために，最適輸送に基づいた単語アラインメントのアプローチを採用する．

3 提案手法

提案手法は，ASR の認識結果のテキストと，知識ベース上のエントリの 2 つのテキストを入力として与える．ASR テキストと各エントリとの間で類似度を計算した後，最も類似度が高かった ASR テキストとエントリのペアを紐付けることで Entity Linking を行う．

以下，単語数 n, n' をそれぞれ持つ 2 つのテキスト（ASR テキスト，エントリ）を $s = \{t_1, t_2, \cdots, t_n\}, s' = \{t'_1, t'_2, \cdots, t'_{n'}\}$ と表す．また，各単語 t_i に対応する単語ベクトルを $w_i \in \mathbb{R}^D$ と表す．

3.1 単語アラインメント

本研究では，単語アラインメントに使用する最適輸送アルゴリズムとして Unbalanced Optimal Transport (UOT)[6][7] を用いる．UOT は，文の長さが大きく異なるテキスト間の類似度を適切に計算できる [5] ため，正式名称と略称との間の類似度を高くできると考えられる．UOT を用いたテキスト s, s' 間の距離の計算式を以下に示す．

$$D(s, s') = \min_{T \in \mathbb{R}_+^{n \times n'}} \sum_{i,j} T_{i,j} d(t_i, t'_j) + \mathrm{reg} \cdot \Omega(T) \\ + \mathrm{reg}_{\mathrm{m1}} \cdot \mathrm{KL}(T^\mathsf{T} \mathbb{1}_n, \mu_s) + \mathrm{reg}_{\mathrm{m2}} \cdot \mathrm{KL}(T \mathbb{1}_{n'}, \mu_{s'}), \tag{1}$$

$$\Omega(T) = \sum_{i,j} T_{i,j} \log(T_{i,j}). \tag{2}$$

ここで，$T_{i,j}$ は i から j への輸送量を，$d(t_i, t'_j)$ は 2 つの単語 t_i, t'_j 間の輸送コスト表している．また，$\Omega(T)$ は Entropic regularization term，$\mathrm{KL}(T^\mathsf{T} \mathbb{1}_n, \mu_s)$ は分布 $T^\mathsf{T} \mathbb{1}_n, \mu_s$ 間の Kullback-Leibler divergence である．$\mathrm{reg}, \mathrm{reg}_{\mathrm{m1}}, \mathrm{reg}_{\mathrm{m2}}$ は各ペナルティ項の影響を調整するハイパーパラメータである．$\mu_s, \mu_{s'}$ はそれぞれ輸送前と輸送後の確率分布を表しており，以下のように表される．

$$\mu_s = \{(t_i, m_i)\}_{i=1}^{n}, \mu_{s'} = \{(t'_j, m'_j)\}_{j=1}^{n'}, \tag{3}$$

$$\sum_i m_i = 1, \sum_j m'_j = 1. \tag{4}$$

ここで，$m_i \in [0, 1]$ は確率質量であり，μ_s は各単語 t_i について質量 m_i の荷物があることを意味している．

3.2 単語間の輸送コスト

WMD では単語間の輸送コストとしてユークリッド距離が用いられているが，ユークリッド距離にはノルム（単語重要度）と偏角（意味的類似度）が混在している [8]．そのため，重要度が大きく異なる単語間では，意味的に近いにもかかわらず，類似度が不当に低く見積もられてしまう恐れがある．そこで，本研究では，横井らの研究 [8] に従い，単語ベクトル間のコサイン距離を輸送コストの計算に使用する．さらに，単語の意味的な類似度だけでなく音声の類似度も考慮するために，単語の音素列間のレーベンシュタイン距離も導入する．2 つの単語 t_i, t'_j 間の輸送コスト $d(t_i, t'_j)$ を以下のように定義する．

$$d(t_i, t'_j) = \lambda \cdot d_{\cos}(w_i, w'_j) + (1 - \lambda) \cdot d_{\mathrm{edit}}(p_i, p'_j). \tag{5}$$

ここで，$d_{\cos}(w_i, w'_j)$ は単語ベクトル w_i, w'_j 間のコサイン距離，$d_{\mathrm{edit}}(p_i, p'_j)$ は単語 t_i, t'_j にそれぞれ対応する音素列 p_i, p'_j の間の正規化レーベンシュタイン距離である．また，$\lambda \in [0, 1]$ は，輸送コスト計算時に意味的類似度をどの程度考慮するかを決定するハイパーパラメータである．

3.3 単語の重み付け

本研究では，知識ベース内における単語の重要度を考慮するために，Inverse Document Frequency(IDF) を確率質量に導入する．しかし，通常の IDF を利用した場合，音声認識誤りの影響で IDF 辞書内の単語とマッチせず適切な単語スコアが計算できなくなる恐れがある．そこで，Soft TFIDF[9] を参考に新たに Phonetic Soft IDF (PS-IDF) を提案する．PS-IDF は音素列の類似度を考慮した IDF であり，IDF 辞書内の単語と完全一致していない場合でも類似した発音の単語から IDF スコアを計算する．PS-IDF の定義を以下に示す．

$$\mathrm{PSIDF}(t_i) = \frac{1}{|\mathfrak{D}(p_i, P^e, \theta)|} \sum_{p_j^e \in \mathfrak{D}(p_i, P^e, \theta)} \mathrm{IDF}(p_j^e) \cdot (1 - d_{\mathrm{edit}}(p_i, p_j^e)). \tag{6}$$

ここで，p_j^e は知識ベース上にあるエントリを構成する単語 t_j^e の音素列，P^e は知識ベース上にあるエントリの単語音素列の集合である．また，$\mathfrak{D}(p_i, P^e, \theta)$ は P^e に含まれる単語音素列 p_j^e のうち，音素列 p_i との正規化レーベンシュタイン距離が $d_{\mathrm{edit}}(p_i, p_j^e) \leq \theta$ であるものの集合であり，θ はしきい値（ハイパーパラメータ）である．

最終的に，2 つのテキストの確率分布は以下のように表される．

$$\mu_s = \left\{ \left(t_i, \frac{\mathrm{PSIDF}(t_i)}{Z} \right) \right\}_{i=1}^{n}, \mu_{s'} = \left\{ \left(t'_j, \frac{\mathrm{PSIDF}(t'_j)}{Z'} \right) \right\}_{j=1}^{n'}. \tag{7}$$

ここで，Z, Z' は正規化定数である．

4 実験と考察

4.1 データセット

提案手法を評価するためのデータセットとして，自社で運用している自動音声対話サービス AI Messenger Voicebot[2] の Entity 辞書および発話ログデータを使用した．このデータセットには，飲食系と医療系の 2 種類のドメインのデータが含まれている．

Entity 辞書とは，商品名や店舗名などのエントリと，そのエントリの同義語を人手で登録したドメイン固有の知識ベースである．エントリの同義語には，エントリの略称や表記ゆれなどのフレーズの他に，音声認識誤りを考慮したフレーズ（e.g. エントリ：Abema Towers，同義語：阿部タワー）も登録されている．Entity 辞書に含まれているエントリ数と同義語数を表 2 に示す．

発話ログデータは，Entity 辞書内のエントリと正しく紐付けられたユーザ発話を収集したものである．データセット内のユーザ発話の件数を表 3 に示す．なお，発話ログデータには同じ内容の発話データが多数含まれているため，実際に収集された発話データの件数を「重複あり」，重複を除いた発話データの件数を「重複なし」で示している．実験では，重複ありデータを用いて実際の運用時の性能を，重複なしデータを用いて発話パターンのカバー率を評価した．

表 2 各 Entity 辞書内のエントリ数と同義語数

	飲食系ドメイン	医療系ドメイン
エントリ数	57	50
同義語数	300	171

表 3 ユーザ発話件数

	飲食系ドメイン	医療系ドメイン
重複あり	571	3445
重複なし	110	208

4.2 実験設定

提案手法を評価するにあたり，テキスト間類似度を測る教師なし手法として，音素列の正規化レーベンシュタイン距離，WMD[3]，WRD[4]，Lazy-EMD[5] を採用した．本実験では，単語分散表現として chiVe[10] の事前学習済み Magnitude モデル (v1.2 mc30)[3] を採用し，形態素解析器には SudachiPy[11][4] を使用した．SudachiPy の分割モードは A モード（UniDic 短単位相当）とした．テキストから音素列への変換は pyopenjtalk[5] を用いた．PS-IDF のしきい値 θ は 0.1 に設定した．UOT を使用している手法（提案手法，Lazy-EMD）の Entropic regularization term の係数 reg は，Lazy-EMD の実験に倣い 0.009 に設定した [5]．また，Kullback-Leibler divergence にかかる 2 つの係数 $\mathrm{reg}_{\mathrm{m1}}, \mathrm{reg}_{\mathrm{m2}}$ は，略称を発話するユーザ側のテキストの重要度を高くするために

2) https://www.ai-messenger.jp/voicebot/

3) https://github.com/WorksApplications/chiVe

4) https://github.com/WorksApplications/SudachiPy

5) https://github.com/r9y9/pyopenjtalk

$reg_{m1} = 1.0, reg_{m2} = 0.5$ に設定した．

4.3 Entity 辞書内実験

提案手法が音声認識誤りフレーズを含む多様な同義語に対処できるのかを評価するために，Entity 辞書に登録された同義語を入力として与え，その同義語に対応するエントリを紐付ける Entity 辞書内実験を行った．このとき，紐付け対象であるエントリセットには，Entity 辞書内にある同義語は含まれないものとする．評価結果を表 4 に示す．提案手法は，飲食系，医療系の両方のドメインで比較手法よりも高い性能を示した．

次に，提案手法の各構成要素の効果を確かめるためにアブレーション実験を行った．実験結果を表 5 に示す．ここで，`-Word vector` は，提案手法から単語分散表現の要素を抜いたもの，すなわち輸送コスト計算時に音素列の類似度のみを考慮したモデルである．一方，`-Phone` は単語分散表現の類似度のみを考慮したモデルである．`-PS-IDF` では，単語の重み付けに PS-IDF を利用せず，WMD と同様に一様な重みを付与した．`-UOT` は，最適輸送アルゴリズムに UOT を利用せず，WMD と同様に Earth Mover's Distance(EMD) を用いた．

結果より，飲食系ドメインにおいては提案手法が最も高い性能を示していることが分かる．一方，医療系ドメインでは PS-IDF を利用しないモデルが最も性能が高くなった．医療系ドメインの Entity 辞書の同義語には，飲食系ドメインと比較して音声認識誤りフレーズが多数登録されていたことから，PS-IDF が音声認識誤りにうまく対処できず，適切な単語の重みが計算できなかったことが要因と考えられる．

表 4 Entity 辞書内実験の結果（Accuracy[%]）．

	飲食系ドメイン	医療系ドメイン
Levenshtein	69.0	60.8
WMD[3]	69.0	59.6
WRD[4]	64.7	57.9
Lazy-EMD[5]	74.0	60.2
Proposed	**77.0**	**70.2**

4.4 発話ログデータを用いた実験

次に，実際のユーザ発話が与えられた際の Entity Linking の性能を評価した．評価結果を表 6, 表 7 に示す．提案手法は，飲食系ドメインにおいては重複あり，重複なしのデータの両方で最も高い性能を示した．一方，医療系ドメインにおいては，重複ありのデータではレーベンシュタイン距離が最も高い性能を示した．提案手法の性能が低くなった原因として，ユーザが頻繁に言及していた特定のエントリについて適切に紐付けできなかったことが挙げられる．ただし，提案手法は重複なしのデータでは最も高い性能を示しているため，比較手法よりもエントリや発話の多様なパターンをカバーしていると考えられる．

表 5 アブレーション実験結果（Accuracy[%]）．

	飲食系ドメイン	医療系ドメイン
- Word vector	75.7	74.9
- Phone	72.7	55.6
- PS-IDF	76.7	**76.0**
- UOT	74.7	68.4
Proposed	**77.0**	70.2

表 6 発話ログデータを用いた評価実験の結果（重複あり，Accuracy[%]）

	飲食系ドメイン	医療系ドメイン
Levenshtein	84.4	**93.4**
WMD[3]	84.8	86.7
WRD[4]	84.2	82.9
Lazy-EMD[5]	89.0	91.3
Proposed	**89.1**	91.9

表 7 発話ログデータを用いた評価実験の結果（重複なし，Accuracy[%]）

	飲食系ドメイン	医療系ドメイン
Levenshtein	64.5	70.2
WMD[3]	68.2	67.8
WRD[4]	67.3	66.3
Lazy-EMD[5]	75.5	72.1
Proposed	**76.4**	**80.3**

5 まとめ

本研究では，意味的類似性と音韻的類似性の両方を考慮した単語アラインメントに基づく教師なし Entity Linking 手法を提案した．比較実験の結果，提案手法は飲食系ドメインでは最も高い性能を示した．一方，医療系ドメインのユーザ発話データを用いた実験では，提案手法がレーベンシュタイン距離よりも性能が低いという結果が示されたが，多様な発話パターンのカバー率という面では比較手法よりも優れていることが示唆された．今後は，より効果的な音韻的特徴を組み込んだ Entity Linking 手法の検討を進めていきたい．

参考文献

[1] Yong Zhao, Jinyu Li, Shixiong Zhang, Liping Chen, and Yifan Gong. Domain and speaker adaptation for cortana speech recognition. In **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 5984–5988, 2018.

[2] Arushi Raghuvanshi, Vijay Ramakrishnan, Varsha Embar, Lucien Carroll, and Karthik Raghunathan. Entity resolution for noisy ASR transcripts. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations**, pp. 61–66, 2019.

[3] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In **Proceedings of the 32nd International Conference on International Conference on Machine Learning**, Vol. 37, p. 957–966, 2015.

[4] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator's distance. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2944–2960, 2020.

[5] Yimeng Chen, Yanyan Lan, Ruinbin Xiong, Liang Pang, Zhiming Ma, and Xueqi Cheng. Evaluating natural language generation via unbalanced optimal transport. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20**, pp. 3730–3736, 2020.

[6] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. **Mathematics of Computation**, Vol. 87, No. 314, pp. 2563–2609, February 2018.

[7] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. In **Proceedings of the 28th International Conference on Neural Information Processing Systems**, Vol. 2, p. 2053–2061, 2015.

[8] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator's distance. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2944–2960, 2020.

[9] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In **Proceedings of the 2003 International Conference on Information Integration on the Web**, p. 73–78, 2003.

[10] 真鍋陽俊, 岡照晃, 海川祥毅, 一馬, 内田佳孝, 浅原正幸. 複数粒度の分割結果に基づく日本語単語分散表現. 言語処理学会第 25 回年次大会 (NLP2019), pp. NLP2019–P8–5, 2019.

[11] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.