ラベル付き系列予測による音声シグナルの Textless 依存構造解析

神藤駿介　宮尾祐介
東京大学
{skando,yusuke}@is.s.u-tokyo.ac.jp

概要

音声言語処理技術の多くは音声認識モデルとテキスト処理モデルを直列接続することで実現されるが，近年は音声認識を介さずに（=Textless に）音声表現ベクトルから直接的に学習を行う研究も盛んである．本研究ではその中でも特に依存構造解析に注目し，音声表現ベクトルを用いた Textless 構文解析の可能性を探る．提案手法は，音声表現ベクトルから依存構造を表すラベル付き系列を直接予測する Textless なアプローチを取る．直列的なアプローチを取る既存手法との比較実験の結果，Textless な依存構造解析は十分可能である一方，頻度の低い依存関係の予測においては語彙情報の寄与が大きいことが示唆された．

1　はじめに

近年，音声言語処理に対する新しいアプローチとして Textless NLP[1)] という試みがある [1, 2, 3]．これは，音声認識モデルとテキスト処理モデルを直列接続する従来のアプローチと異なり，陽に音声認識を行わずに音声表現ベクトルを直接用いて学習を行うというものである．Textless なアプローチを取ることで，テキスト資源の少ない言語のモデリングを促進したり，音声に含まれるニュアンスや感情の情報を効果的に利用できるなどといった効果が見込まれる．また，乳児はテキストを使わず音声のみから言語に関する多くの知識を獲得できることを考えると，Textless NLP は認知言語学や発達心理学への寄与もあり得る研究分野である [4]．

本研究では Textless NLP の取り組みのうち，特に依存構造解析に注目する．依存構造解析は，単語間の依存関係を予測して構文構造を出力する技術であり，自然言語処理において長年研究されている基礎技術の一つである．音声シグナルから依存構造解析を行う研究として wav2tree [5] があるが，これは従来式の直列接続によるアプローチとなっている．本研究では，図 1 に示すように，音声表現ベクトルから依存構造を表すラベル付き系列を直接予測する Textless なアプローチを提案する．同じ枠組みで POS タグを予測する研究はあるが [6]，単語間の依存関係，すなわち主辞の位置や依存関係のラベルを予測することができるかどうかは未知である．

実験の結果，Textless な依存構造解析は十分可能である一方，頻度の低い依存関係の予測においては語彙情報の寄与が大きいことが示唆された．これは言語の二重分節性から来る帰結であるとも考えられる．今後の方向性として，どのような音声言語処理タスクが Textless に実行可能なのかを調査することで，人間の言語獲得過程に関する知見の提供も期待される．

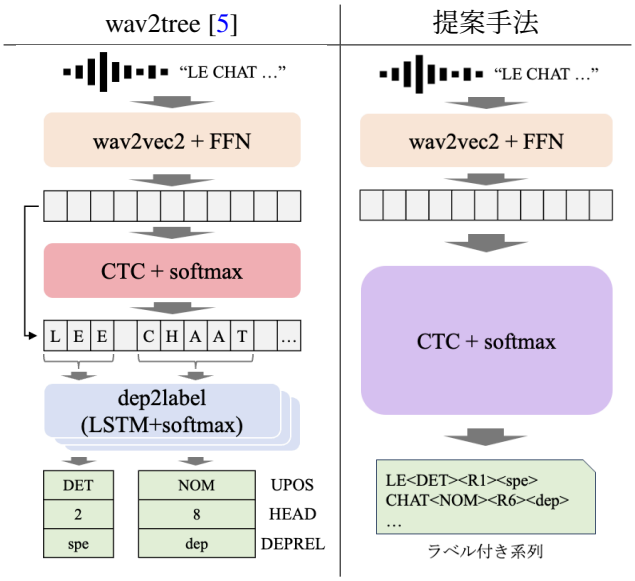


図 1　既存手法 (wav2tree [5]) と提案手法の比較．既存手法は音声認識モジュール (CTC+softmax) を挟む一方，提案手法は音声表現ベクトルから依存構造を表すラベル付き系列を直接予測する Textless なアプローチを取る．

1)　https://speechbot.github.io/

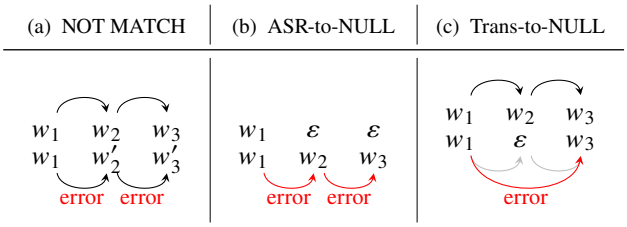


図 2　木構造の書き換え規則．上側が正解 (Transcription) の依存構造を，下側が音声認識結果の単語列と書き換え後の依存構造を表す．書き換え規則の適用後に追加されるアノテーションを赤色・消去されるアノテーションを灰色で示した．

2　先行研究: wav2tree

我々の知る限り，wav2tree [5] は音声シグナルを入力として End-to-End に構文解析を行う唯一の研究である．本研究は wav2tree をベースラインとして同じ設定で実験を行うため，本節で詳細に説明する．

2.1　モデルの構成

図 1 の左で表されているように，wav2tree は以下の三つのモジュールからなる：

1. 表現抽出モジュール (wav2vec2+FFN)
2. 音声認識モジュール (CTC+softmax)
3. 構文解析モジュール (dep2label)

1 は wav2vec2 [7] に 3 層の全結合層を加えたものとなっている．2 は全結合層と softmax 関数からなり，CTC loss [8] によって学習する．予測対象の語彙は文字の集合と空白文字からなり，空白文字の間の区間が一単語とみなされる．3 では，まず CTC のデコード結果をもとに各単語の区間を定め，音声表現ベクトルの該当区間を単語埋め込みとみなす．それらを dep2label [9] に入力して依存構造解析を行う．dep2label は各単語の UPOS・HEAD の相対位置・DEPREL（依存関係ラベル）を予測することで依存構造解析を行うモデルである．

2.2　音声認識エラーへの対応方法

wav2tree は依存構造解析の前に音声認識を行うため，音声認識エラーがあった場合に正解の構文木との乖離が生まれる．そこで，訓練時に正解の構文木を音声認識エラーに応じて書き換える処理が挟まる．具体的な手順は以下の 2 ステップである．

1. 正解・予測単語列のアラインメントをとる[2)]
2. Yoshikawa ら [10] の規則で木構造を書き換える

2 の規則は図 2 で示すように以下の 3 つからなる：

(a) NOT MATCH: アラインメントは取れるが認識エラーがあるケース ($w_2 \neq w'_2, w_3 \neq w'_3$)．該当単語の関係ラベルを error に書き換える．

(b) ASR-to-NULL: 音声認識結果に過剰に単語が存在するケース．該当単語の直前の単語を主辞とし，関係ラベルを error とする．

(c) Trans-to-NULL: 音声認識結果に単語が不足するケース．該当単語に接続するエッジを消去する．該当単語を主辞とする単語が存在した場合 (w_3)，該当単語の主辞に付け替えた上で error ラベルを付ける．

2.3　結果のまとめと課題

筆者らは wav2tree と同じ構成で End-to-End ではない（=各モジュールを独立に学習させる）手法をベースラインとして比較実験を行った．実験の結果，wav2tree はベースラインに比べて構文解析の精度が高く，End-to-End に音声から構文解析を行うことは可能であると結論づけられている．

wav2tree は音声認識と構文解析のモジュールが分離しているが，それによって生じる弊害もある．2.2 節で説明した通り，wav2tree は訓練時に音声認識エラーに対応して訓練データ自体を書き換える必要があり，望ましくない処理であると言える．また，訓練データの書き換えがオーバーヘッドとなって訓練に多大な時間がかかることも指摘されている．

3　提案手法: Textless 依存構造解析

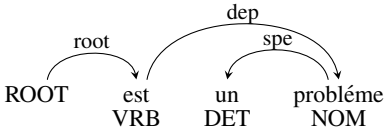
図 1 の右に提案手法の構成を示した．wav2tree との違いは，音声認識モジュールを間に挟まない Textless なアプローチとなっていることである．これは音声特徴ベクトルから 3.1 節で定義される**ラベル付き系列**を直接予測することによって実現される．本手法は，予測系列に音素と POS タグの情報を付与して音声認識を行う先行研究 [6] から着想を得たものである．

3.1　ラベル付き系列の定義

文 $w_1, w_2, \ldots w_n$ の各単語 w_i に，依存構造を表す以下のアノテーションが付いているとする[3)]：

2)　sclite などといった専用のツールを用いる:
https://github.com/usnistgov/SCTK

3)　依存構造のアノテーションは CoNLL-U format とする:
https://universaldependencies.org/format.html




```
est<POS1><L1><REL0>_
un<POS2><R1><REL2>_
probléme<POS0><L2><REL1>
```

図 3 依存構造とそれに対応するラベル付き系列の例．スペース位置を_で明示した．

- p_i：Universal POS タグ (UPOS)
- h_i：主辞のインデックス ($h_i \in [0, n]$)
- r_i：依存関係ラベル

root である単語の主辞のインデックスは 0 とする．

この依存構造を表すラベル付き系列は，以下で定義される系列を $i = 1$ から n まで空白区切りで連結したものとして定義される：

$$w_i\text{<Enc}_{\text{upos}}(p_i)\text{><Enc}_{\text{head}}(h_i - i)\text{><Enc}_{\text{rel}}(r_i)\text{>}$$

ここで，各 Enc はラベルのエンコーダーであり，以下のように定義される．

$$\text{Enc}_{\text{upos}}(p) = \text{POS}\{id(p)\}$$

$$\text{Enc}_{\text{head}}(j) = \begin{cases} \text{L}\{-j\} & (j < 0) \\ \text{R}\{j\} & (J > 0) \end{cases}$$

$$\text{Enc}_{\text{rel}}(r) = \text{REL}\{id(r)\}$$

図 3 にラベル付き系列の例を示す．以降，読みやすさのため，UPOS と依存関係のラベルは適宜エンコードせずに <VRB> や <dep> などと表記する．

3.2 ラベル付き系列のデコード方法

予測系列から依存構造へデコードする際には，まず予測系列をスペースで分割してトークン列 $t_1, t_2, \cdots, t_n$ を得る．各 t_i に対し以下のように w_i, p_i, h_i, r_i を定める．

1. 一番左にある “<” の直前までを w_i とする
2. 一番左にある “<POSj>” をデコードして p_i，“<Rj>” または “<Lj>” をデコードして h_i，“<RELj>” をデコードして r_i とする
3. 2 でラベルが存在しなかった場合，p_i = X，h_i = None，r_i = dep とする

ここで，h_i = None は主辞が存在しないことを意味する．本研究では wav2tree [5] と同様に依存構造の制約として (1)root の一意性 (2) 主辞の一意性 (3) 非巡回性の 3 つを与えるため，[9] で用いられているヒューリスティクスによって後処理を施し，これらの制約を担保する．

3.3 提案手法のポイント

提案手法は，音声認識モジュールを挟まない Textless なアプローチを取ることで，2.3 節で指摘した既存手法の課題（訓練データの書き換え，訓練時間の増大）を解決できる．

一方で，**単語表現ベクトルではなく，音声表現ベクトルから直接的に単語間の依存関係を学習できるのか**は未知であり，本研究で答えたい主要な問いである．

4 実験設定

モデルの各パラメタは wav2tree とほぼ同様だが，以下の点が異なる：

1. 全結合層でバッチ正規化層の代わりにレイヤー正規化層を用いる．
2. 予測対象の語彙を文字集合の代わりに語彙数 1000 の Byte Pair Encoding (BPE, [11]) とする．ただし，各ラベルは User Defined Symbols として登録する．

1 の理由は，事前実験で最高精度を記録したためである．2 の理由は，wav2tree に比べて予測対象の系列が長くなるため，CTC の制約に収まるようにトークン長を出来るだけ減らすためである．

評価は Orféo Treebank [12] で行い，音声認識の指標として WER, CER を，構文解析の指標として UPOS の精度, UAS, LAS を測定する．依存構造の評価においては 2.2 節で説明した木構造の書き換えを行う．wav2tree の評価結果は，筆者らが公開しているモデル[4] を用いて得る．

5 結果と議論

5.1 単語間の依存関係の学習可能性

実験結果を表 1 に示す．UAS, LAS に注目すると，提案手法の性能は wav2tree に僅かに劣るものの，その差はそれほど大きくないと言える．このことから，音声表現ベクトルから Textless に単語間の依存関係を学習することは可能であると考えられる．

4) https://gricad-gitlab.univ-grenoble-alpes.fr/pupiera/wav2tree_release/-/tree/main#pretrained-models

表 1　実験結果．音声認識指標は低いほど良く，構文解析指標は高いほど良い．平均訓練時間は，10 epoch 訓練したときの 1 epoch あたりの平均値．

	音声認識指標		構文解析指標			モデルの性質の比較	
	WER	CER	UPOS	UAS	LAS	パラメタ数	平均訓練時間
wav2tree	30.5	18.2	75.4	70.2	66.0	350M+33M	2.8 h
提案手法	28.4	19.3	77.2	68.6	64.5	350M	1.3 h

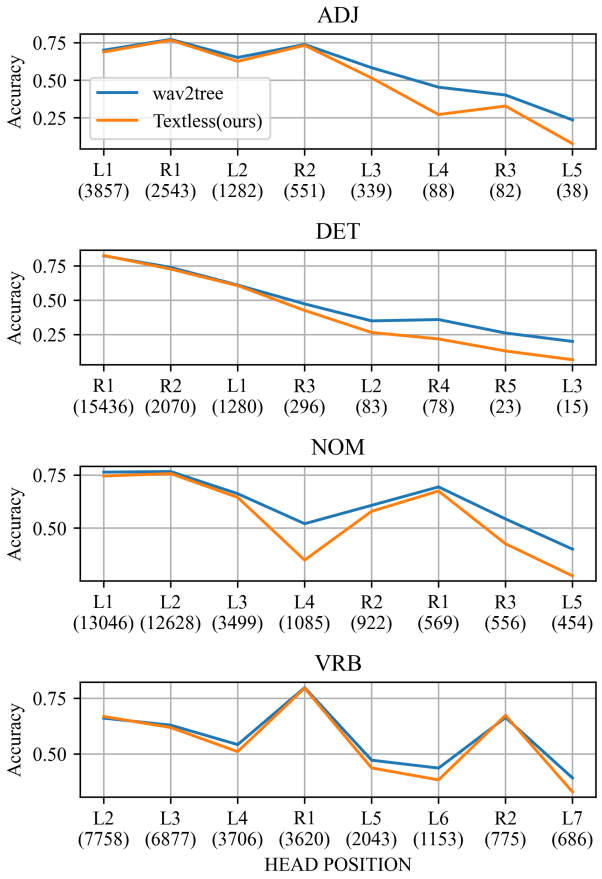


図 4　主辞位置の頻度ごとの予測精度．括弧内の数値は各 POS タグと主辞位置の共起頻度である．頻度上位 8 位までの結果を示す．

分析: 主辞位置の頻度ごとの予測精度　この主張の適用範囲を探るため，主辞位置の頻度ごとの予測精度を比較する．分析のモチベーションは，単語や POS タグに対応する主辞の相対位置や依存関係はある程度決まってしまうケースも考えられることである．例えば，フランス語の定冠詞 “le” は基本的に直後の名詞が主辞・依存関係は spe (specifier) と決まるので，予測系列は “le<DET><R1><spe>” となるケースがほとんどである．したがって，より頻度の少ない系列に対しどれほど頑健に予測できているかを測定することで，単語間依存関係の学習能力をより正確に定量化できる．

ここでは，4 つの代表的な POS タグ (ADJ, DET, NOM, VRB) に対する主辞の相対位置ラベルの頻度ごとに予測精度を比較した．図 4 に結果を示す．総じて wav2tree の方が低頻度な位置予測に対して頑健で，頻度の低い依存関係の予測においては語彙情報の寄与が大きいことが示唆されている．また，比較的近い依存関係（R1, R2）では提案手法も高い精度で予測できることが観察された．

5.2　その他の指標に関する議論

「音声認識指標」について，提案手法の音声認識性能は落ちていないことが分かる．提案手法は単語列だけでなくラベルを同時に予測するモデルとなっているため，これは良い結果であると言える．ただし，wav2tree は予測対象の語彙が文字の集合である一方，提案手法は BPE を用いているため，この条件を揃えた追実験が必要である．

「構文解析指標」の UPOS は提案手法が wav2tree を上回っており，[6] の主張と一貫している．

「モデルの性質の比較」から見て取れるように，提案手法はより少ないパラメタ・より少ない訓練時間で学習できており，3.3 節で述べた期待通りの結果が得られている．

6　おわりに

本研究では，音声表現ベクトルを用いた Textless な依存構造解析の可能性を探った．実験の結果，Textless な依存構造解析は十分可能である一方，頻度の低い依存関係の予測においては語彙情報の寄与が大きいことが示唆された．

この結論は言語の二重分節性から来る帰結であるとも言える．今後どのような音声言語処理タスクが Textless に実行可能なのかを調査することで，語彙の情報を必要とするタスクとそうでないタスクとの境界が浮き彫りになるのではないだろうか．この研究を通して，人間の言語獲得過程に関する知見の提供も期待される．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2237_4 の支援を受けたものです.

参考文献

[1] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1336–1354, 2021.

[2] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 860–872, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 250–266, 2023.

[4] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. **Cognition**, Vol. 173, pp. 43–59.

[5] Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jerome Goulian. End-to-End Dependency Parsing of Spoken French. In **Interspeech 2022**, pp. 1816–1820. ISCA.

[6] Motoi Omachi, Yuya Fujita, Shinji Watanabe, and Matthew Wiesner. End-to-end ASR to jointly predict transcriptions and linguistic annotations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1861–1871. Association for Computational Linguistics.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In **Proceedings of the 23rd International Conference on Machine Learning**, ICML '06, p. 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[9] Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. Viable Dependency Parsing as Sequence Labeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 717–723. Association for Computational Linguistics.

[10] Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. Joint Transition-based Dependency Parsing and Disfluency Detection for Automatic Speech Recognition Texts. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1036–1041. Association for Computational Linguistics.

[11] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[12] Christophe Benzitoun, Debaisieux jeanne marie, and Henri-José Deulofeu. Le projet orfÉo : un corpus d' étude pour le français contemporainthe orfeo project: a study corpus for contemporary french. **Corpus**, 10 2016.