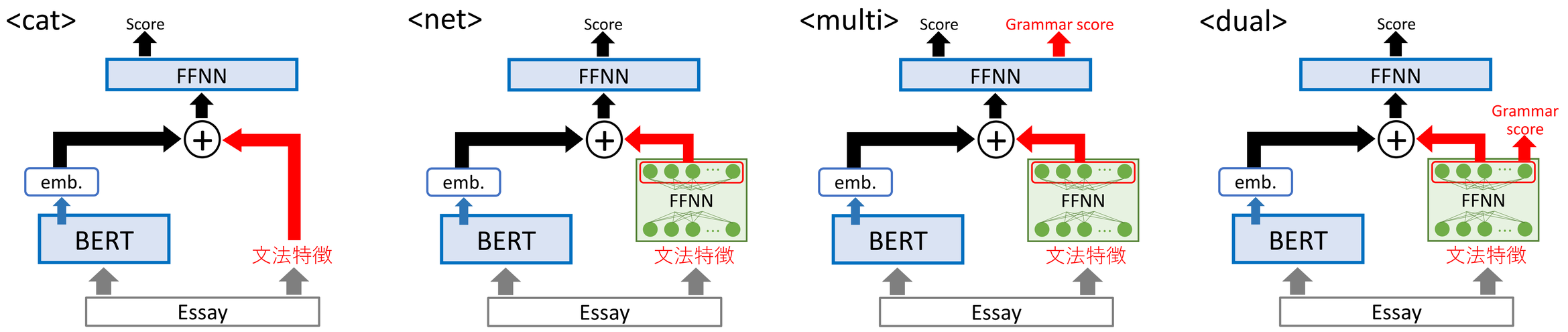
[P4-18]
文法項目の多様性と誤り情報を利用したエッセイ自動採点

○土肥 康輔，須藤 克仁，中村 哲 (NAIST)

本研究の概要

エッセイの分散表現＋**文法特徴** で自動採点

- **正しく使えている文法**と**文法誤り**の情報をモデルの入力に使用 → モデル性能が向上
- エッセイの総合スコアと文法スコアで**マルチタスク学習** → モデル性能がさらに向上



上部のFFNN
隠れ層の数: cat {1, 2, 3, 4, 5, 7, 10}
net {1, 2, 3}
multi, dual {2, 3}
ノード数: 512

文法特徴のFFNN
隠れ数の数 = 3
ノード数 = 文法特徴の次元数の1/2

multiのFFNNはタスク固有層なし
multi, dualでの主タスクの重み {0.8, 0.6}
バッチサイズ {4, 8, 16, 32} を探索

文法特徴

- CEFR × **基準特性** [Hawkins+ 2012]
  - あるレベルの学習者に特徴的な文法項目が存在
  - 人間の採点者はこれらの項目を探しながら評価

特徴量

type256

- CEFR-J Grammar Profile 文法項目頻度分析プログラム [石井+ 2020] を用いて抽出

err24, err54

- GECtoR-large [Tarnavskyi+ 2020] で文法誤りを訂正
- ERRANTの誤りタグと誤りタイプに基づき集計

データ

- ASAP, ASAP++ [Mathias+ 2018]
  - 8つのエッセイ課題に関する答案
  - 総合スコア＋観点別スコア

課題	エッセイ数	スコア範囲
1	1,783	2-12
2	1,800	1-6, 1-4
3	1,726	0-3
4	1,772	0-3
5	1,805	0-4
6	1,800	0-4
7	1,569	0-30
8	723	0-60

- 5分割交差検証で評価
  - [Taghipour+ 2016]の分割
  - train/dev/test = 60/20/20
- 課題ごとに独立して評価
- 評価指標
  - 2次の重み付きカッパ係数 (QWK)

実験

モデル構造による比較 (type256, QWK test)

Model	課題番号 1	2	3	4	5	6	7	8	avg.
baseline	.807	.659	.671	.805	.799	.803	.819	.749	.764
+ type256									
cat	.818	.675	.673	.819	.806	.808	.832	.734	.771
net	**.822**	**.684**	.685	.811	.804	.813	.834	.746	.775
multi	.812	.682	.694	.817	.808	**.814**	**.837**	.749	.777
dual	.820	.675	**.700**	**.820**	**.809**	.806	.831	**.763**	**.778**

- 全てのモデル構造でベースラインよりも向上
- dualのモデル構造で最も高いスコアを達成

文法特徴量による比較 (dual, QWK test)

Features	課題番号 1	2	3	4	5	6	7	8	avg.
baseline	.807	.659	.671	.805	.799	.803	.819	.749	.764
type256	.820	.675	.700	.820	.809	.806	.831	.763	**.778**
err24	.814	.666	.690	.817	.809	.805	.833	.761	.774
err54	.820	.682	.677	.820	.803	.823	.836	.758	.777
type256 + err24	.821	.674	.686	.827	.799	.809	.833	.751	.775
type256 + err54	.821	.679	.690	.816	.816	.810	.835	.754	**.778**
Yang+ 2020	.817	.719	.698	.845	**.841**	**.847**	.839	.744	**.794**

- 詳細な誤り情報を使うほうが高いスコアを達成
- type256とerr24/err54を組み合わせても相乗効果なし

—議論—
- type256をCEFR-Jレベルごとに集計した特徴量
  - W&I+L, EFCamDat, FCEでは効果なし [土肥+ 2023]
  - ASAPではスコアが向上する場合あり

→ 複数のエッセイ課題に基づく答案が含まれているデータセットでは×

- 既存モデルに対しては.01〜.015程度のビハインド

→ 課題：特徴量のより効果的な組み合わせ方の検討