サブワード系列の変化が固有表現抽出に与える影響の調査

辻 航平 [1]　平岡 達也 [2]　鄭 育昌 [1,2]　岩倉 友哉 [1,2]
[1] 奈良先端科学技術大学院大学　[2] 富士通株式会社
tusji.kohei.tl1@naist.ac.jp
{hiraoka.tatsuya,cheng.yuchang,iwakura.tomoya}@fujitsu.com

概要

サブワード正則化によって，複数の分割結果を考慮することで，翻訳や文書分類タスクで精度の向上が得られている．しかし，複数のサブワード系列を用いた推論の調査はあまり行われてはいない．本稿では，固有表現抽出において，複数のサブワード系列から得られた推論結果と本来の推論結果との関係を調査した結果を報告する．結果として，サブワード分割に依らずに一致した推論結果を多く得られていた文はそうでない文よりも精度が良いことが確認でき，モデル間で比較した場合，サブワード分割に依らない推論ができているモデルほど精度が上がることが確認できた．また，固有表現抽出においてもサブワード正則化を用いた単一モデルアンサンブルにより精度を向上させる余地があることも分かった．

1　はじめに

サブワード分割 [1, 2, 3] は近年の自然言語処理で広く利用されている手法であり，単語をより細かいサブワードの列として扱うことで未知語や低頻度語の抑制，辞書数の制限などが行える．しかし，サブワード分割は学習時のデータに対して統計的に処理するため，推論時に出現する単語のサブワード系列が学習時に多く学習したサブワードの列になるとは限らない．

そこで，複数のサブワード分割を用いるサブワード正則化が研究されている．サブワード正則化は，単一の文章から複数のサブワード系列を生成し学習することで，推論時に新規のサブワードや系列が出現することを低減させる．図 1 を例にすると，通常の学習では Tom / has と分割されてモデルに入力されるところ，Tom / ha / s や T / om / has のようなサブワード分割についても学習を行う．

ユニグラム言語モデルを用いたサブワード正則化 [3] は，n-best のサブワード系列を取得し，学習のイテレーションごとに用いるサブワード系列を変更することで，モデルをよりロバストにするための手法である．しかしながら，ユニグラム言語モデルと異なり，サブワード正則化を用いるために必要な n-best のサブワード候補を尤度に基づいて取得する手法は，BERT [4] などで用いられている wordpiece [1] や，RoBERTa [5] などで用いられているバイトペア符号化（Byte Pair Encoding，以下 BPE） [2] では利用できない．そこで，wordpiece に対しては最長一致候補を単語ごとにランダムに棄却することで複数のサブワード系列を取得する MaxMatch-Dropout [6] が，BPE に対しては結合時にランダムに結合を行わないことで複数のサブワード系列を取得する BPE-Dropout [7] が提案されている．

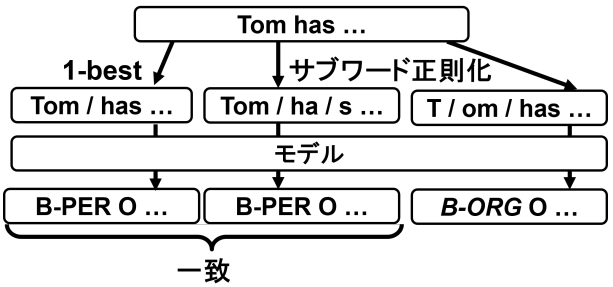
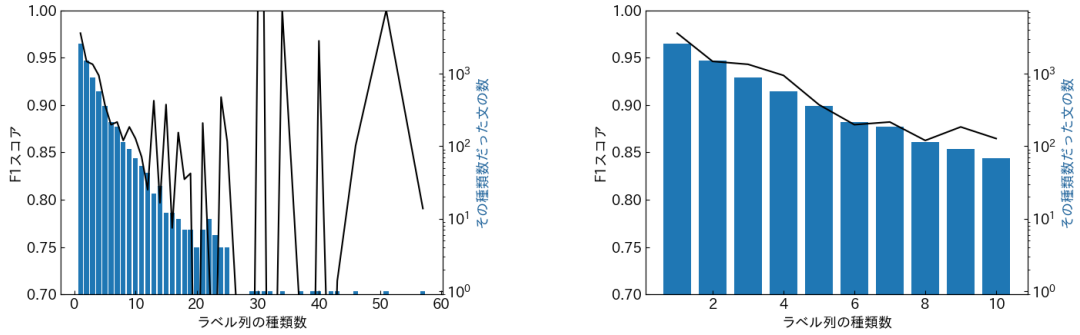


図 1: 複数のサブワード系列を用いて単一モデルから複数の出力を取得する例

また，図 1 のようにサブワード正則化によって複数の出力を得た場合でも，サブワード分割前の文は同一であることから，同一文からの異なる分割結果に対して，予測分布の差を小さくするように学習する手法 [8] や，推論時にもサブワード正則化を用いることで単一モデルから複数の出力を取得し，それらをアンサンブルすることで単一モデルアンサンブルを行う手法 [9] も研究されている．

しかしながら，先行研究では，複数のサブワードを用いた精度改善手法が主であり，異なるサブワードに対する出力への影響調査は深く行われていな



(a) 全データに対しての結果　　**(b)** 10 種類以下に絞った結果

図 2: ラベル列の種類数と 1-best の F1 スコアの関係．折れ線グラフが F1 スコア，ヒストグラムが文の数を表している．

い．本稿では，サブワード正則化によって得られる複数のサブワード系列を同一モデルに入力した際の推論結果に対して調査を行うことで，サブワード系列の変化と推論結果の関係の調査と今後の精度改善に向けた議論を行う．

先行研究 [3, 7, 6, 9] では，サブワード正則化は，ニューラル機械翻訳タスク，文書分類タスクといった生成，分類タスクに対して用いられてきた．しかしながら，本稿では，サブワード系列の変化と推論結果の関係を明らかにする目的から，抽出タスクである固有表現抽出（Named Entity Recognition, 以下 NER）が適していると考え，調査対象とする．

2 実験設定

2.1 データセット

CoNLL-2003 [10] および，CoNLL-2003 と同様の定義で 2020 年の記事で新しく作成されたテスト用データセット CoNLL++ [11] を用いた．学習には CoNLL-2003 の訓練データを，評価には，CoNLL-2003 の開発用データおよび評価データ，CoNLL++ を用いた．

2.2 推論方法

本稿ではサブワード系列の変化が推論に与える影響を調べるため，推論時にサブワード正則化を適用する．

推論には，CoNLL-2003 の訓練データで fine-tuning した $\text{BERT}_{\text{BASE}}$ [4], $\text{BERT}_{\text{LARGE}}$ [4], $\text{RoBERTa}_{\text{BASE}}$ [5], $\text{RoBERTa}_{\text{LARGE}}$ [5]，$\text{LUKE}_{\text{LARGE}}$ [12] を用いた．このとき，LUKE 以外のモデルはエポック数 20，バッチサイズ 32，学習率 5e-5 ($\text{RoBERTa}_{\text{LARGE}}$ のみ 1e-5)，weight decay 0.01 で fine-tuning した．LUKE については Hugging Face 上に公開されている fine-tuning 済みモデル[1)]を利用した．また，LUKE 以外の 4 モデルに対しては同様のハイパーパラメータでサブワード正則化を用いて fine-tuning したものも用いた．このとき，サブワード正則化は，BERT に対しては MaxMatch-Dropout (MMD) [6] を，RoBERTa には BPE-Dropout (BD) [7] をそれぞれ利用し，ハイパーパラメータ棄却率 p = 0.1 および 0.3 で学習した．

学習したモデルを用いてテストデータそれぞれに対してサブワード正則化なしの推論 1 回とサブワード正則化ありの推論 100 回の計 101 回の推論を行った．このとき，サブワード正則化なしの推論 1 回を 1-best とする．

2.3 評価方法

各文で得られた 101 個の推論で出現したラベル列の種類数と，1-best の文ごとの F1 スコアの関係を調査した．ラベル列の種類数について図 1 を例に説明すると，3 種類のサブワード系列を入力に 3 つの出力を得たとき，そのうち 2 つが同じ B-PER O... というラベル列で，残りの 1 つが B-ORG O... というラベル列だったため，この文のラベル列の種類数は 2 種類になる．

F1 スコアは seqeval[2)]を用いて計算した．seqeval において正解している O ラベルは計算から省かれる．本調査では 1 文ごとの F1 スコアを計算するため，正解ラベルと推論ラベルがどちらもすべて O ラベルの場合があり，このとき seqeval では F1 スコアが得られない．この場合，調査対象から除いた．

1) https://huggingface.co/studio-ousia/luke-large-finetuned-conll-2003
2) https://github.com/chakki-works/seqeval

表 1: 各モデルの各データセットでの 1-best の文ごとの F1 スコアとラベル列の平均種類数および，101 個の推論がすべて一致していた割合（一致率）．vanilla はサブワード正則化なしで学習したモデル．

モデル		CoNLL-2003 検証			CoNLL-2003 テスト			CoNLL++		
		F1	平均種類数	一致率	F1	平均種類数	一致率	F1	平均種類数	一致率
$\text{BERT}_{\text{BASE}}$	vanilla	95.13	3.61	0.36	91.30	3.42	0.39	84.65	7.10	0.11
	+MMD 0.1	95.89	1.44	0.8	91.43	1.63	0.75	82.83	3.41	0.37
	+MMD 0.3	96.19	1.35	0.84	92.12	1.48	0.8	81.86	3.03	0.44
$\text{BERT}_{\text{LARGE}}$	vanilla	95.96	3.11	0.38	92.40	2.91	0.41	87.17	6.04	0.12
	+MMD 0.1	96.39	1.41	0.81	92.46	1.55	0.77	84.95	3.12	0.41
	+MMD 0.3	96.44	1.29	0.86	92.65	1.39	0.82	86.45	2.78	0.46
$\text{RoBERTa}_{\text{BASE}}$	vanilla	96.91	3.85	0.38	92.81	3.63	0.43	92.81	7.20	0.12
	+BD 0.1	96.74	3.40	0.42	92.79	3.29	0.45	92.95	6.05	0.13
	+BD 0.3	96.65	1.33	0.84	92.86	1.47	0.82	91.72	2.65	0.49
$\text{RoBERTa}_{\text{LARGE}}$	vanilla	97.15	2.79	0.41	93.21	2.73	0.43	94.61	4.35	0.22
	+BD 0.1	97.15	1.31	0.85	93.22	1.42	0.82	94.23	2.00	0.61
	+BD 0.3	96.77	1.23	0.88	92.94	1.30	0.86	92.74	0.86	0.65
$\text{LUKE}_{\text{LARGE}}$	vanilla	96.88	2.25	0.54	94.20	2.11	0.58	94.81	3.82	0.30

3 実験結果

3.1 種類数と F1 スコアの関係

図 2 に $\text{RoBERTa}_{\text{LARGE}}$ モデルでの実験結果を示す．どちらの図もラベル列の種類数に対しての，それぞれその種類数だった文の 1-best の平均 F1 スコア（折れ線グラフ）とその種類数だった文の数（ヒストグラム）である．左図はすべてデータセットに対する結果であり，種類数が多くなると，文の数も減り，平均 F1 スコアが安定していない．そこで，観測数がそれなりに得られる種類数が 10 種類以下を対象とした調査も行った．図 2 の右図から，ラベル列の種類数が 10 種類以下の範囲では，種類数が少ないほど，つまり，一致した出力結果が得られているほど，平均 F1 スコアが上がっていることが確認できる．

3.2 モデル間の評価

3.1 節で単一モデル内で異なる文を比較した際に推論で表れたラベル列の種類数が少ないほど平均 F1 スコアが上がっていることが確認できた．本節では異なるモデル間で比較をすることで，一致した出力と精度の関係を調査する．

それぞれのモデルに対して，各データセット全体での 1-best の F1 スコアとラベル列の平均種類数および，101 個の推論がすべて一致していた割合（ラベル列の種類数が 1 だったデータの割合）を示したものが表 1 である．

表 1 から，ほとんどの場合でサブワード正則化による精度の変化は，モデルの違いによる精度の変化よりも小さいの対して，サブワード正則化を用いた推論は用いていない推論よりも一致率が大きく上昇し，平均種類数は大きく減少していることがわかる．

3.3 アンサンブルモデル

翻訳タスクにおいて，サブワード正則化によって得られた複数の推論結果をアンサンブルすることで精度が向上したことが報告されている [9]．そこで，実験で取得した推論 101 個を用いて，最も多く推論されたラベルを選択する多数決によるアンサンブルに加えて，複数推論結果を用いた際に出せる精度の上限を調べるために，ラベル単位で最も良い結果になるように選択した上界を表 2 に示す．

表 2 の多数決の結果では一部精度向上がみられ，また，上界の結果からより高い精度が得られる見込みも得られた．

4 考察

4.1 ラベル数と F1 スコアの関係

3.1 節では出力ラベル列の種類数が少ないほど F1 スコアが高くなりやすいことを報告した．ここで，ラベル列の種類数は 1 文のラベル数が少ないほど，取れる種類数が減るため少なくなりやすい．このとき，ラベル数が少ない文の F1 スコアが高いと，ラベル列の種類数と平均 F1 スコアは疑似相関となってしまう．そこで，各文のラベル数と F1 スコアの関係を調べたものが図 3 になる．図の直線は回帰直線であり，その傾きがわずかに正であることから，ラベル数が少ないほどが F1 スコアが高くなりやす

表 2: 各モデルでの単独モデルアンサンブルの結果 (F1 値)．括弧内は 1-best との差．太字は 1-best より精度が改善しているもの．

モデル		CoNLL-2003 検証 多数決	CoNLL-2003 検証 上界	CoNLL-2003 テスト 多数決	CoNLL-2003 テスト 上界	CoNLL++ 多数決	CoNLL++ 上界
$\text{BERT}_{\text{BASE}}$	vanilla	**95.14 (0.01)**	**98.09 (2.96)**	**91.31 (0.01)**	**95.82 (4.52)**	**85.05 (0.40)**	**93.35 (8.70)**
	+MMD 0.1	**96.01 (0.12)**	**97.93 (2.04)**	**91.71 (0.28)**	**95.09 (3.66)**	**82.94 (0.11)**	**91.31 (8.48)**
	+MMD 0.3	**96.37 (0.18)**	**97.96 (1.77)**	92.04 (-0.08)	**94.96 (2.84)**	81.71(-0.15)	**89.91 (8.05)**
$\text{BERT}_{\text{LARGE}}$	vanilla	**95.98 (0.02)**	**98.35 (2.39)**	**92.50 (0.10)**	**96.21 (3.81)**	87.01 (-0.16)	**93.41 (6.24)**
	+MMD 0.1	96.16 (-0.23)	**98.18 (1.79)**	92.32 (-0.14)	**95.24 (2.78)**	84.73 (-0.22)	**92.19 (7.24)**
	+MMD 0.3	**96.46 (0.02)**	**97.98 (1.54)**	92.62 (-0.03)	**94.70 (2.05)**	**86.81 (0.36)**	**92.68 (6.23)**
$\text{RoBERTa}_{\text{BASE}}$	vanilla	96.78 (-0.13)	**99.13 (2.22)**	92.62 (-0.19)	**96.62 (3.81)**	**93.16 (0.35)**	**97.19 (4.38)**
	+BD 0.1	96.72 (-0.02)	**99.11 (2.37)**	92.73 (-0.06)	**96.71 (3.92)**	92.58 (-0.37)	**97.44 (4.49)**
	+BD 0.3	**96.76 (0.11)**	**98.28 (1.63)**	**92.94 (0.08)**	**95.22 (2.36)**	**91.94 (0.22)**	**95.69 (3.97)**
$\text{RoBERTa}_{\text{LARGE}}$	vanilla	**97.21 (0.06)**	**98.91 (1.76)**	**94.01 (0.80)**	**96.39 (3.18)**	94.21 (-0.40)	**97.26 (2.65)**
	+BD 0.1	**97.27 (0.12)**	**98.53 (1.38)**	**93.33 (0.11)**	**95.67 (2.45)**	93.87 (-0.36)	**96.85 (2.65)**
	+BD 0.3	**96.80 (0.03)**	**98.48 (1.71)**	92.70 (-0.19)	**94.67 (1.73)**	92.67 (-0.07)	**95.43 (2.69)**
$\text{LUKE}_{\text{LARGE}}$	vanilla	96.72 (-0.16)	**98.38 (1.50)**	**94.21 (0.01)**	**96.74 (2.54)**	**94.98 (0.17)**	**97.41 (2.60)**

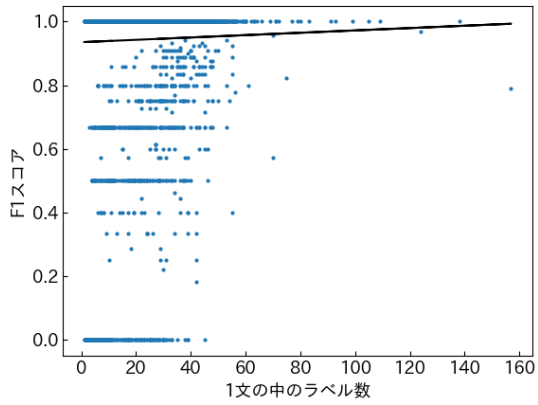


図 3: 1 文におけるラベルの数とその文の $\text{RoBERTa}_{\text{LARGE}}$ モデルの 1-best の F1 スコアの関係

いとは言えない．

4.2 アンサンブルモデル

3.3 節では単一モデルアンサンブルによる精度を示した．多数決法では最大で+0.40 ポイントの改善がみられ，合計で 22 件が改善しているものの，ほとんどが ±0.2 ポイント以内に収まっており，残りの 17 件では低下していた．さらに，上界の結果ではどのモデルも少なくとも 1.5 ポイントは上昇できる可能性があるため，複数のサブワード系列を用いた推論は NER において精度を向上させる余地があることがわかった．

また，ほとんどのモデルにおいて，サブワード正則化のハイパーパラメータを上げるほど上界のスコアが減少している．これは表 1 の一致率からわかるように，サブワード正則化によってほとんどの出力が同じになっているため，サブワード正則化なしの場合に少数のサブワード系列でのみ得られていた推論結果が得られなくなっているためだと思われる．

4.3 サブワード正則化での学習

3.2 節の結果より，サブワード正則化を行うことで通常よりも多くの文に対して一致した出力を得ることは可能になるものの，モデルの精度はあまり改善していない．しかし，サブワード正則化を行っていないモデル同士を比べた際には，ほとんどの場合で一致率が高いモデルのほうが精度が高くなっている．そのため，現在のサブワード正則化とは異なる手法を用いて一致した出力の取得を目指すことで，NER の精度を上げることができる可能性がある．

5 おわりに

NER において，サブワード分割に依らない予測結果が得られた場合，その結果が正しい可能性は，得られなかった場合よりも高いことが示唆された．モデル間で比較を行った場合，サブワード分割に依らない推論ができているモデルほど精度が上がることが確認できたが，サブワード正則化を用いて学習することでサブワード分割に依らない推論結果を得た場合には精度の改善はほとんど見られないことも示された．また，単純な多数決でも精度が改善する場合も多く，上界の結果より，複数のサブワード系列を用いることで精度を向上させる余地があることも分かった．これらの結果から，今後は既存のサブワード正則化とは異なる方法で，サブワード分割に依らない推論ができる手法や，多数決よりも良いアンサンブル手法について研究を行っていく．

参考文献

[1] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast WordPiece tokenization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2089–2103, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. **CoRR**, Vol. abs/1907.11692, , 2019.

[6] Tatsuya Hiraoka. MaxMatch-dropout: Subword regularization for WordPiece. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 4864–4872, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[7] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1882–1892, Online, July 2020. Association for Computational Linguistics.

[8] Xinyi Wang, Sebastian Ruder, and Graham Neubig. Multi-view subword regularization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 473–482, Online, June 2021. Association for Computational Linguistics.

[9] Sho Takase, Tatsuya Hiraoka, and Naoaki Okazaki. Single model ensemble for subword regularized models in low-resource machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2536–2541, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[11] Shuheng Liu and Alan Ritter. Do CoNLL-2003 named entity taggers still work well in 2023? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8254–8271, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6442–6454, Online, November 2020. Association for Computational Linguistics.