トークン長に着目した汎用言語モデル利用拡大のための一考察

齋藤 彰　　山崎 貴宏　　村田 稔樹
沖電気工業株式会社 経営基盤本部 研究開発センター
{saitou659, yamasaki635, murata656}@oki.com

1 はじめに

BERT や ELMo に代表される汎用言語モデルの活用により，読解や文書分類など様々な分野で精度向上が報告されている [1][2]．一方で，汎用言語モデルを直接適用できない場合がある．例えば，入力テキストの一部分を抽出するときは，汎用言語モデルのトークンでは適切な出力を構成できない可能性がある．このように，汎用言語モデルを利用する場合そこで採用されるトークンが後処理にとって必ずしも最適であるとは限らない．そのため，一般に公開されている汎用言語モデルを利用する場合，利用者にとっての最適なトークンが使えなくなってしまう．

この課題を解決するためには，特定のトークン単位で解析した結果を，他のトークン単位の解析結果に変換する技術が効果的であると考える．本論文では，これをトークン長変換タスクとして新しく提案する．また，系列ラベリングを使って本タスクを試行した結果も合わせて示す．

2 汎用言語モデルの利活用における課題

汎用言語モデルは，事前学習に使用したトークン化手法により生成するトークン列を入力とし，そのトークン列ごとに解析結果を出力する．そのため，汎用言語モデルの利用者は，そのトークン列ごとに生成される解析結果に合わせて後処理を構成する場合が一般的である．しかしながら，事前学習に使用したトークン化手法により生成するトークン列は，後処理において必ずしも最適であるとは限らない．その例として，以下のような入力文の一部の文字列が解答となる場合について考える．

目的 重要単語抽出

入力文 梅干しはクエン酸を含む

解答 梅干し，クエン酸，含む

ここで，事前学習の際に使用したトークン化手法を SentencePiece[3] とし，SentencePiece によりトークン化した以下のトークン列を汎用言語モデルへの入力に用いたとする．

トークン列 梅/干/し/は/ク/エン/酸/を含む

このトークン列は，後処理で必要な単語列とは異なるものである．このとき，汎用言語モデルの解析結果を直接適用する場合，後処理ではこのトークン列の要素を一単位とするため，解答の文字列をトークン列から容易には構成できない．

本論文では，このような課題を解決するために，特定のトークン化手法により生成したトークン列を異なるトークン長のトークン列に変換する技術が効果的であると考え，その処理をトークン長変換タスクとして新たに提案する．

3 トークン長変換タスク

トークン化されたテキストを他のトークン長へ変換可能か検証する系列ラベリング問題を提案する．本論文では，以下のような基底トークン列 T_{in} から目的トークン列 T_{out} を推定することをトークン長変換タスクと呼び，課題として設定する．

T_{in} ギ/ター/の/穴/を/サ/ウン/ド/ホール/と/呼/ぶ/。

T_{out} ギター/の/穴/を/サウンドホール/と/呼ぶ/。

このタスクでは，T_{out} のトークンそれぞれについて，T_{in} から同じトークン長で推定できたかを判定する．上述の例に対する推定結果が次の $\widehat{T_{out}}$ である場合，T_{out} と同じトークン長で推定できたトークンは「ギター」,「と」,「呼ぶ」,「。」の 4 つであるため，それらを正答とする．

$\widehat{T_{out}}$ ギター/の穴を/サウンド/ホール/と/呼ぶ/。

4 実験

4.1 トークン長変換タスクのデータセット

Wikipedia 日本語版のテキストを 1 文ごとに SentencePiece[3] と mecab-ipadic-NEologd[4] を使用した MeCab[5] のそれぞれでトークン化し，一方を基底トークン列，他の一方を目的トークン列とすることで，トークン長変換タスクのデータセットを作成した．データセットの用途別の内訳は，表 1 の通りである．

表 1: データセットの内訳

学習用	開発用	評価用
5,000 文	500 文	500 文

4.2 検証した手法

系列ラベリングに有効とされる CRF，LSTM-CRF，BiLSTM-CRF の 3 つの手法を用いて，トークン長変換タスクに有効な手法を検証した．

検証した手法では図 1 に示すように，基底トークン列のトークンそれぞれの文字列と連続性を表現する IOB2 タグを入力の素性とし，目的トークン列のトークンの連続性を表現する IOB2 タグを推定することにより，トークン長変換タスクを系列ラベリング問題として解いた．

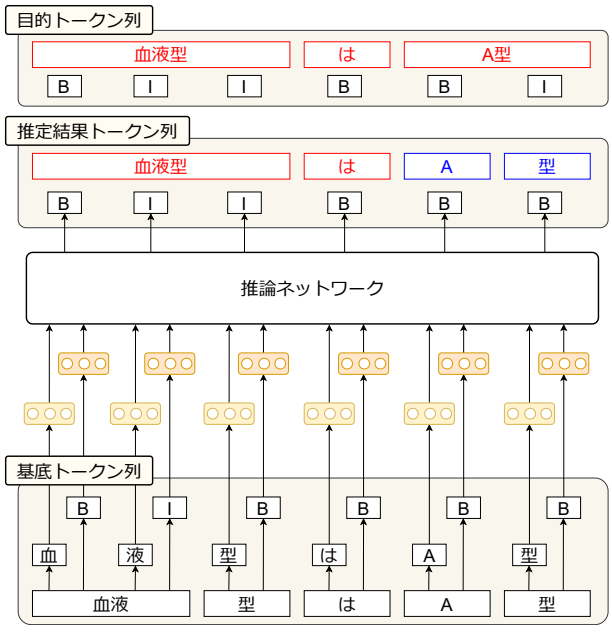

図 1: ネットワーク構成

5 評価結果

トークン単位の評価結果を表 2，IOB2 タグによる文字単位の評価結果を表 3 示す．

表 2: トークン単位の評価結果

	適合率	再現率	F 値
CRF	0.501	0.479	0.489
LSTM-CRF	0.824	0.807	0.816
BiLSTM-CRF	**0.856**	**0.838**	**0.847**

表 3: 文字単位の評価結果

	タグ	適合率	再現率	F 値
CRF	B	0.793	0.759	0.776
	I	0.810	0.828	0.819
	O	0.890	**1.000**	0.942
LSTM-CRF	B	0.935	0.916	0.925
	I	0.931	0.944	0.937
	O	**0.977**	0.995	**0.985**
BiLSTM-CRF	B	**0.950**	**0.930**	**0.940**
	I	**0.942**	**0.956**	**0.949**
	O	0.972	0.997	0.984

このとき，表 2 と表 3 の結果から，次の 2 点が言える．

1. LSTM-CRF と BiLSTM-CRF が CRF の精度を上回った
2. 文字単位の精度に比べて，トークン単位の精度が低い

第 1 に，LSTM-CRF と BiLSTM-CRF は，トークン単位の F 値において 0.30 以上 CRF を上回っていたことが挙げられる．このことから，CRF に入力の順序を考慮して学習可能な LSTM または BiLSTM が加わることにより，複数の入力を考慮した識別が可能になったためと考えられる．また，BiLSTM-CRF と LSTM-CRF のトークン単位の F 値を比較すると，BiLSTM-CRF が僅かに上回る結果であった．第 2 に，いずれの構成においても，トークン単位の F 値は文字単位の IOB2 タグの F 値に比べて低いことが挙げられる．この理由として，正解トークンと推定結果でトークン長が異なる場合が含まれていたと考えられる．その分析結果については，6 章で述べる．

6 考察

5 章の評価結果を受けて，トークン長ごとの精度を調査した結果を表 4 に示す． 表 4 の結果から，トーク

表 4: トークン長ごとの評価結果

	トークン長	トークン数	適合率	再現率	F 値
CRF	all	9,061	0.501	0.479	0.489
	1 文字	3,948	0.633	0.589	0.610
	2 文字	2,941	0.573	0.496	0.532
	3 文字以上	2,172	0.216	0.239	0.227
LSTM-CRF	all	9,061	0.824	0.807	0.816
	1 文字	3,948	**0.930**	0.893	**0.911**
	2 文字	2,941	**0.879**	0.791	0.833
	3 文字以上	2,172	0.655	0.675	0.665
BiLSTM-CRF	all	9,061	**0.856**	**0.838**	**0.847**
	1 文字	3,948	0.907	**0.915**	**0.911**
	2 文字	2,941	0.869	**0.841**	**0.855**
	3 文字以上	2,172	**0.670**	**0.676**	**0.673**

表 5: 学習データ量を拡充した場合の評価結果

	トークン長	トークン数	適合率	再現率	F 値
CRF	all	408,947	0.554	0.533	0.543
	1 文字	191,358	0.669	0.641	0.655
	2 文字	126,592	0.629	0.556	0.590
	3 文字以上	90,997	0.254	0.274	0.264
LSTM-CRF	all	408,947	0.926	0.913	0.919
	1 文字	191,358	0.967	0.947	0.957
	2 文字	126,592	0.935	0.912	0.923
	3 文字以上	90,997	0.831	0.841	0.836
BiLSTM-CRF	all	408,947	**0.934**	**0.924**	**0.929**
	1 文字	191,358	**0.972**	**0.953**	**0.962**
	2 文字	126,592	**0.942**	**0.925**	**0.933**
	3 文字以上	90,997	**0.847**	**0.859**	**0.853**

ン長が大きくなるほど精度が低くなっていることが分かる．この理由として，トークン長 3 文字以上のトークンが学習用データに不足していたことが考えられる．

そのため，データセットのデータ量を学習用 200,000 文，開発用 20,000 文，評価用 20,000 文に引き上げ，追実験を行った．その追実験におけるトークンごとの評価結果を表 5 に示す．

表 4 と表 5 の比較からデータ量を大きくした場合，いずれの構成においても精度が向上していることが分かる．そのため，学習用データの増加により，精度向上が期待できる．

一方で，学習用データの拡張を行わず，機械学習による精度を上げる方法については，次の 2 点が挙げられる．

1. 汎用言語モデルを活用する
2. トークンの表現に用いるタグを拡張する

第 1 の方法は，機械学習のネットワーク構成に汎用言語モデルを組み込み，入力トークンの解析能力向上を目指すものである．汎用言語モデルを入力トークンの解析に活用することにより，入力テキストに含まれる単語の出現頻度や順序を考慮して，トークン長の推定が可能になると期待できる．ただし，4.2 で述べたネットワーク構成は文字単位で処理するため，組み込む汎用言語モデルの扱うトークン長を文字単位とするなどの措置が必要である．

第 2 の方法は，トークンの境界の表現方法を IOB2 タグから BIOES タグなど情報量の多いものに変更することで，トークンの境界の識別能力向上を目指すものである．BIOES タグを活用すると，1 つのタグに隣接可能なタグの制限が強化されるため，隣接するタグの推定が容易になる．ここで，上述の学習の効率を上げる 2 点の方法は未検証のため，今後の課題とする．

7　おわりに

本論文では，トークン長の違いにより汎用言語モデルを直接活用できない場合があることを課題として捉え，その課題を解消するためのタスクとして新たにトークン長変換タスクを提案した．さらに，そのタスクを系列ラベリングにより試行し，5,000 文のテキストの学習により最大 F 値 0.847 でトークン長の変換が可能であることを確認した．

今後は，汎用言語モデルの活用とトークンの表現に用いるタグの拡張について評価実験を行い，それらの有効性について評価する．

参考文献

[1] 田中裕隆, 曹類, 白静, 馬ブン, 新納浩幸. BERT による単語埋め込み表現列を用いた文書分類. 研究報告自然言語処理 (NL), Vol. 2019-NL-240, No. 16, pp. 1–5, 2019.

[2] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-2018)*, Vol. 1, pp. 2227–2237, 2018.

[3] 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 (EMNLP-2018)*, pp. 66–71, 2018.

[4] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP-2017), pp. 875–878, 2017.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237, 2004.