BERT-NER によるテンプレート分類と Patent-Tokenizer を用いた和文特許文章の文章生成

高根沢　光輔†　　堀江 和正‡　　天笠 俊之‡

筑波大学情報学群情報科学類†　　計算科学研究センター‡　　計算科学研究センター‡

要旨

本論文では，特許文章の自動生成を行うシステムを目指し，これを実行するためのトークナイザーを提案する．特許文書はボキャブラリーが特殊で未知語が多いという特徴があるため，一般的なトークナイザーではトークンが短くなりすぎるといった問題があった．本研究では，特許文章を Unigram で学習したトークナイザーと Jumanpp という形態素解析ツールを利用し，できるだけ長い文字列を１トークンとするアイデアを導入することで，適切な長さのトークンを生成できると考えた．このアイデアを導入した新しいトークナイザーPatent-Tokenizer を入力途中の特許文章の次文予測に適用したところ，存在しない単語やフレーズが従来手法に比べ生成されにくくなることが確認できた．

1. 序論

近年，事前学習済み汎用言語モデルの自然言語処理性能が飛躍的に向上している[1]．特に一般文章の生成では，人が作成した文章と見分けがつかないほど自然な文章が生成できる[2].

一方で，特殊な文章を生成するようなタスクでは，汎用モデルが適さないケースがある．例えば特許文書が挙げられる．特許文書は，汎用モデルが目指す文章の自然さよりも，文章の正しさが重視される.ここの正しさとは，実在しない単語（物質名等）を出力しないなどが挙げられる．

既存モデルで特許文章を生成した結果を表１に示す．既存汎用モデル GPT-2 ではもっともらしい文章を生成するが，その内容は特許の適用範囲となる成分列挙となっていない．また，許文章に特化させた LSTM 文章生成モデルであっても，「燥関係熱的薬剤」のような特許文書に存在しない成分名を生成してしまうという問題があった.

表 1 従来手法における入力文章の次文予測の例

入力文章：点眼剤には、他の薬効成分
生成文章（GPT2）：点眼剤には、他の薬効成分だけでなく、様々な合成着色料が入っています。その合成着色料は、アレルギー体質の人にはあまり適していない成分も多く、毎日何年も使用すると、眼の皮膚や健康に支障が出てしまうことも。
生成文章(既存 LSTM)：点眼剤には、他の薬効成分、例えば、乾燥関係熱的薬剤（大型又は、た形成分配合成分）である移動物、通して測定のもの技術が移動通して高速度グリドを作用途への３の技機能などのビールオーのレゾ化合成分子から、ポリンなどの…
正解文章：点眼剤には、他の薬効成分、例えばグリチルリチン酸二カリウム、イプシロンアミノカプロン酸、アラントイン、アズレンスルホン酸ナトリウム、硫酸亜鉛などの抗炎症剤…

こういった事象は，モデルの学習状況に加えて，トークナイザーに問題があることが示唆されている[3].

トークナイザーの役割は，自然文を単語列に分割し，数値計算可能な単語 ID に変換することである．GPT-2 1B では，モデルやトークナイザーが汎用的なフレーズを学習しているため，未知語よりも自然文が高確率で生成されてしまう問題がある．これにより，成分名の列挙よりも自然な文章生成が主な生成文になるような挙動に至る．また，一般的なトークナイザーでは，短い文字列をトークンとして抽出してしまい，これが存在しない単語の生成といった問題に繋がっている．

このような問題に対して，本論文では，適切に表現できない特許文章を効率よく表現できる Patent-Tokenizer を提案することで，特許文章の文章生成精度の向上を目指す．また，Patent-Tokenizer による特許文章向けトークナイザーの効果を最大化するため，特許文章の分野判定を行う BERT-NER モデルを組み合わせた LSTM モデルを構築した.

関連研究については，英文向けトークナイザーが獲得するボキャブラリーの内容を改善することで，テキスト分類タスクの精度を向上させる研究事例がある[3]．この事例では，トークナイザーが出力する単語列を正しい単語の位置で区切るように調整することで，既存 NLP モデルの精度向上を行ったものである．

本論文では，和文特許文章向けのトークナイザーを提案し，特許文書における文章生成精度の向上を目指す．また，特許文書における多様な分野ごとに特化モデル生成し，文章生成時に利用するためのテンプレート判定を行う BERT-NER モデルを作成した．評価実験では，提案手法 Patent-Tokenizer を用いて構築した特化モデルと汎用モデルの違いについても考察する．

貢献：既存 Tokenizer では未知語を含む成分名の表現が困難であった事を踏まえ，Tokenizer を改善することで未知語を含む成分名を生成しやすくする Patent-Tokenizer を提案する．提案手法では，既存手法よりも特許文章のトークン表現が単語基準でトークン化されやすくなることを確認した．また，特許文章の分野別の文章スタイルを考慮するため，BERT-NER によるテンプレート判定と，判定結果に基づいて分類した分野特化文章生成モデルを組み合わせて利用することで，既存手法に Patent-Tokenizer を利用した場合よりもさらに精度向上することが確認できた．

2. Patent-Tokenizer

Patent-Tokenizer は，既存のトークナイザーに用いられている SentencePiece Unigram Tokenizer と，日本語の分かち書きが行える Jumanpp を組み合わせて実装されている．

SentencePiece Unigram Tokenizer では，入力する特許文章を分割し，高頻度に出現する文字列範囲を 1 トークンとして抽出する．特許文章を入力として与えることにより，未知語を含む成分名を高くトークナイズできることが期待される．

次に，Jumanpp については，分かち書きにより入力文章を単語ごとに分割することができる．既知の日本語単語列を Unigram Tokenizer よりも高精度に単語分割することができるため，既知語のトークナイズ精度が高いことが期待される．

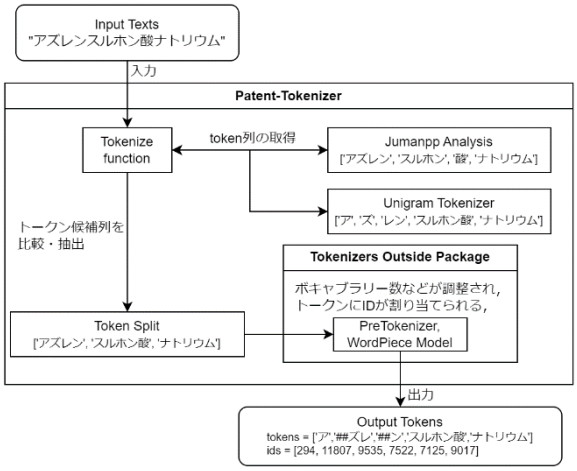

図 1 Patent-Tokenizer の全体アルゴリズム

Patent-Tokenizer では，この 2 つの列の要素を先頭から読み出し，1 トークンに含まれる文字数が多い方を出力トークンとする形をとる．また，入力文字と出力トークンの辻褄が合うようにするため，次に結合させる文字列は，前に採用したトークンから次に採用するトークンまでの間の文字列とする．これにより，Unigram Tokenizer や Jumanpp で既知語と抽出できた文字列は最長一致で抽出され，その間にある未知語文字列も範囲選択で漏れなくトークン抽出できる．

Tokenizer の動作例として，特許文書で未知語を含むような成分名を含む「グリチルリチン酸二カリウム、イプシロンアミノカプロン酸、アラントイン、アズレンスルホン酸ナトリウム」という文をトークナイズした．既存手法である GPT-2 1B が利用する T5-Tokenizer，提案手法である Patent-Tokenizer のそれぞれのトークナイズ結果を比較したものを表 2 に示す．

表 2 では，既存手法の T5-Tokenizer ではトークン数が 28 トークンに分割されていることがわかる．対して，Patent-Tokenizer では 14 トークンと分割数が少なくなっていることがわかる．これは 1 つの成分単語名を構成する際に少ないトークンで表現することが可能となる．1 トークンに未知語を含む成分名を多く含めることで，文章生成時に未知語の生成を容易にすることが期待される．

表 2: 既存手法と提案手法のトークナイズ結果.

Tokenizer	トークナイズ結果
T5	['グリ','チル','リ','チン','酸','二','カリウム','、','イ','プ','シ','ロン','アミノ','カ','プロ','ン','酸','、','ア','ラント','イン','、','アズ','レン','スル','ホン','酸','ナトリウム'](28 tokens)
PatentT	['グリチルリチン', '酸', '二', 'カリウム', '、', 'イプシロン', 'アミノ', 'カプロン酸', '、', 'アラントイン', '、', 'アズレン', 'スルホン酸', 'ナトリウム'] (14 tokens)

3. BERT-NER

特許文書は，文章中に多様な単語を利用している．これらの多くは分野ごとに利用される単語名が異なり，単語が定まれば分野が定まる関係を持つ．本研究ではこの特徴に着目し，成分名をカテゴリー名と見てその成分名の周辺語を学習させた特化モデルを作成した．なお，BERT-NER モデルは BERT モデルを NER 用にファインチューニングさせている．ファインチューニング元となる BERT モデルは，特許文章２４万文（約７０００万文字）をゼロから学習させたオリジナルの BERT モデルを利用した．

また，周辺語の学習データを BERT-NER モデルに学習させることで，周辺語からカテゴリー名となる成分名を予測する BERT-NER モデルを作成した．この特化判別モデルと特化文章生成モデルを利用することで，分野に応じた適切な文章が生成できることが期待される．

BERT-NER モデルの入力にトークナイズした入力文章のトークン列を与えることで，出力として１トークンごとに最も確率の高いカテゴリー値を返す．このカテゴリー値は前後関係を含めて判定しているため，文章構成によって同じトークンでも判定結果が異なる場合がある．また，１トークンごとにカテゴリー分類を行うため，BERT-NER の出力結果として複数カテゴリー列挙される場合がある．そのような場合には，最も出現回数の多いカテゴリー値を最終的な特化モデルの利用先として決定するようにした．(表 3)

表 3 BERT-NER モデルの入出力

入力トークン	BERT-NER の出力結果
['このように', '、', '帯状', '金具', '１', 'の', '下部', '管', '支持', '枠', '４', 'と', '上部', '管', '支持', '枠', '５', 'と', 'を', '形成', 'する', 'こと', 'で', '、']	['B-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'I-フィルム', 'O', 'O', 'O']

4. 評価

評価実験では，既存モデルと提案手法の文章生成精度実験および，特許文書データベースに存在する単語列がどれだけ生成できているか評価する実験を行った．

評価対象モデルについては，既存モデルは汎用モデル GPT-2 及び，既存手法で特許文章を学習した LSTM 系モデルを３つ用意した．提案手法では，特許文章を Patent-Tokenizer を用いて学習させた LSTM モデル PatentT(Common)と，入力文章からカテゴリー判定と利用モデルを判定する PatentT(BERT-NER)の２つを評価モデルとして利用した．

精度評価実験では，各モデルによる文章生成精度を ROGUE-1, ROGUE-2, ROGUE-L で比較した(表 4)．結果を見ると、ROUGE-1 では、提案手法 Patent-Tokenizer を用いた LSTM モデルが他モデルよりも最も高い精度が出ることを確認した。とくに、BERT-NER を用いて特化モデルの分類と判定を行ったモデルでは、ROUGE-2, ROUGE-L を含めて全体的な精度が改善できることを確認した。

また，従来手法では難しかった特許文章の生成精度の改善に Patent-Tokenizer のようなトークナイザーの改善は文章生成精度の向上に有力であるということを確認した。とくに，特許文章で重要視される，存在する単語を正しく生成するという観点でも，提案手法が最も優れる性能を持つことを確認した．（表５）

表 4 文章生成モデルの精度評価

モデル名	ROGUE-1	ROGUE-2	ROGUE-L
GPT-2 1B	0.294	0.125	0.230
LSTM-char	0.128	0.003	0.086
LSTM-unigram	0.225	0.073	0.170
LSTM-gpt-t5	0.237	0.081	0.180
PatentT(BERT-NER)	**0.380**	**0.212**	**0.322**
PatentT(Common)	0.303	0.049	0.250

次に，生成した文章が特許データベースに含まれる単語列をどれだけ生成できているかどうか含まれる割合を評価したものを示す．

既存モデル GPT-2 や従来手法の学習済み LSTM 系モデルと，提案手法 Patent-Tokenizer を用いた LSTM モデルの生成文章が，学習データの元となっている特許データベースから抽出した単語列にどれだけ含まれているか調査したものを示

す(表 5)．これは，特許文章の特徴である句読点で区切った間の文章が成分名になることが多いという特徴を利用したものである．なお，ここで示す PatentT の評価は，PatentT(BERT-NER)モデルによって生成された文章データをもとに割合を計算している．

表 5　特許データベースの含有率測定結果

モデル名	トークン	実在単語トークン	割合
GPT-2 1B	885	424	47.9%
LSTM-char	939	340	8.5%
LSTM-unigram	920	326	35.4%
LSTM-gpt-t5	939	340	36.2%
PatentT	**1167**	**655**	**56.1%**

最後に，BERT-NER の学習時の状況と精度について示す．正解ラベルを正しく予測できた精度としては 37%ほどであり，学習データ，テストデータの範囲では一定学習できていることを確認した．

BERT-NER モデルの学習状態を示したもの．分類対象は 344 種類あり，学習時の Train-Loss, Validation-Loss 及び，テストデータの正解ラベル付き文章を正しく分類した精度を示す(表 6)．

表 6 BERT-NER の学習モデル

Categories	Train-Loss	Valid-Loss	Acc
344 types	0.257	0.706	37.4%

5. 考察

第一に，文章生成が正解文章とどれくらい同じ文章を生成したか評価する ROGUE では，Patent-Tokenizer による文章生成モデルの学習と，BERT-NER による特化モデルの判定と利用特化モデルの決定を行う方式が最も精度が高い結果を得た．とくに，ROUGE-2 については，Patent-Tokenizer で汎用文章生成モデルを学習した PatentT[Common]では 0.049，特化モデルを分野別に学習させ，分野判定を行う PatentT[BERT-NER]モデルでは 0.212 となっており，大きな差が確認できた．これは，連続する文字列が正解データにより近い結果を文章生成モデルが生成したことを意味しており，未知語を含むような特許文章の複雑な文書構造をうまく学習することができていることを示している．今回の PatentT[Common]と PatentT[BERT-NER]は一つの文章生成モデルのサイズが同じである．そのため BERT-NER による分野別の特化モデルを用意することで，学習パラメーターがより最適化できたのではないかと考察することができる．

また，追加で特許データベースに含まれる未知語を含むフレーズがどれだけ含まれているか評価実験を行った．結果としては，既存手法よりも多くの特許文章フレーズ（成分名など）を生成していることを確認することができた．このことから，Patent-Tokenizer を改善することで，未知語を含む特許フレーズを生成しやすくチューニングできるということが今回の実験で示すことができた．

6. 結論

Patent-Tokenizer をベースに文章生成モデルを構築することで，既存汎用モデルのファインチューニングでは達成できない，未知語への対応が可能となり，未知語を含む特許文章に特化した文章が生成されることを確認した．また，純粋な ROGUE による精度評価についても，既存手法よりも優れた結果を得ることができた．提案手法の評価実験では，Patent-Tokenizer を使って特許文章全体を学習させた汎用 LSTM モデルと，分野判定モデル BERT-NER を用いた特化 LSTM モデルを２つ用意した．結果としては，どちらも既存手法を評価指標で上回る結果が得られた．さらに，Patent-Tokenizer に BERT-NER を用いたほうがより正しく，特許単語データベースに含まれるような正しい文章が生成されやすくなることを確認した．

また，GPT-2 1T モデルと LSTM モデルだとモデルサイズが全く異なるため，LSTM のような小さなモデルでも有効な文章生成が行えるモデルが作成可能ということが確認できた．最後に，和文特許文章の文書生成モデルにおいて，トークナイザーの改善が精度向上に有力な手法であることを確認した．

[1] Tokenizer の違いによる性能が変わる事例 Tokenizer の違いによる日本語 BERT モデルの性能評価（築地俊平，新納浩幸,情報処理学会第 27 回年次大会 発表論文集，2021 年 3 月）

[2] All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text (Clark et al., ACL-IJCNLP 2021)

[3] An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers (Hofmann et al., ACL 2022)