試験区分: 情報科学区分
氏名: 夏見 昂樹
希望研究室: 自然言語処理学研究室
現在の専門: データマイニング

1　これまでの修学内容

私は現在，龍谷大学先端理工学部電子情報通信課程に在学しており，データマイニング関連の研究室に所属している．3年次には，Cookpad のレシピデータから手順データの利用食材と調理動作の順序性に着目することでレシピの人気分析を行った．卒業論文では,「
」について着手する予定である．

2　NAIST で取り組みたい研究

2.1　はじめに

奈良先端科学技術大学院大学で取り組みたい研究テーマは「日本語の同音異義語誤り部にラティス構造を用いたニューラル機械翻訳の頑健性向上」である．本稿では，研究テーマの背景・関連研究，提案手法，提案手法の評価について述べる．

2.2　背景・関連研究

近年，NMT モデルは目覚ましい発展を遂げているが，ほとんどの NMT モデルではノイズを含んだ入力は翻訳文に悪影響を与えてしまう [1]. 一般的な入力ノイズの一種として同音異義語ノイズが挙げられる．同音異義語ノイズは私たちが文章を作成時に誤った漢字変換をしてしまった際，音声入力時などに発生する．このようなノイズは入力文が膨大になるにつれ人手での検出は困難になり，翻訳時に理想的な出力を獲得するための障壁となる．

このような問題を解消するために昨今，ノイズを含んだ入力文から正確な翻訳文を獲得する研究が行われている．Qin ら [2] は，中国語から英語の翻訳にて同音異義語ノイズのアノテーションデータが不足していることから，人工的にランダムに単語を選択し，同音異義語に置き換えることで検出器の訓練データを作成した．これを用いて同音異義語検出器を学習させ，入力文中の同音異義語誤り率の高い単語を音節に置き換えることで漢字と音節からなる混合シーケンスを作成可能とした．この混合シーケンスを扱えるように学習した NMT を用いることによって，翻訳性能とノイズに対する頑健性が向上することを示した．

本稿で取り扱う日本語に関する研究では，同音異義語の誤りを検出・訂正する手法において三品ら [3] は，従来の ngram のみの手法に比べ，ngram と確率 LSA を組み合わせることによって精度が改善することを示した．また，藤井ら [4] は BERT と Zero-shot 学習を組み合わせての検出を試みた. この他にも様々な同音異義語誤りを検出・訂正する手法が提案されているが，日本語の同音異義語誤りに頑健な NMT モデルの提案は未だ行われていない．

他方，NMT モデルの性能向上を図る取り組みの一例として，入力シーケンスにラティス構造を取り入れる研究が行われている．しかし，近年 NMT モデルとして非常に高い精度を達成した Transformer では，系列内での各トークンの位置情報を D 次元ベクトル表現に符号化する位置符号化が用いられており [5]，ラティス構造をそのまま入力することは困難である [6]．そこで，Xiao ら [6] はセグメンテーションの多様性が NMT の精度に影響を与えるという仮定のもと，ラティス構造を構成するノードの 1 文字目を絶対位置とし，ラティス構造を単一のシーケンスで表現することで Transformer への入力を可能とした．同時にラティスグラフ同士の距離を適切に扱える self-attention を用いたラティスベースのエンコーダを提案し，機械翻訳タスクにて従来の Transformer エンコーダーに比べ優位性を示した．Li ら [7] は中国語 NER タスクにて，ラティス構造を構成する各トークンに始点と終点の位置を割り当てスパンからなる平坦な構造に変換することで Transformer への入力を可能とし，従来法に比べ性能・効率ともに優れていることを示した．

このようにラティス構造を活用した研究は幅広く行われており，本稿で取り扱う問題の同音異義語誤りについて，誤り箇所にラティス構造を用いることで，他の同音異義語候補を複数持つシーケンスを表現することが可能となる．Li ら [7] のラティス構造を NMT へ入力する手法を踏襲することで，NMT モデルが翻訳文に基づいて複数の候補から適切な語を注視するように学習することが期待できる．

以上より「ラティス構造を含んだ入力シーケンスに対応した Transformer を用いた同音異義語に頑健な日本語から他言語への NMT システム」の提案を行う．

2.3　提案手法

提案手法は (1) 同音異義語誤りの検出，(2) 同音異義語誤り部へのラティス構造の付与，(3)Transformer へのラティス構造の入力・訓練，という 3 つのステップに分けられる．以下でそれぞれの詳細を述べる．

2.3.1　同音異義語誤りの検出

日本語ソース文中の同音異義語誤りを検出する手法について述べる．検出器にはトークナイズの方法として，事前学習済みモデルで文を MeCab を用いて形態素解析を行った後に単語をサブワード分割している BERT base Japanese[1] を用いる．また検出器の訓練データとして毎日新聞の記事データを用いる．まず，訓練データに対して形態素解析器 MeCab を用いて漢字部の読みごとの同音異義語辞書を作成する．次に，先ほど MeCab によって検出された漢字部を作成した同音異義語辞書を用いて，誤った同音異義語で置換することで，擬似的に誤りのある文を作成する．

また，我々が誤った同音異義語を選択してしまう場面は，同じ同音異義語でも文中での出現頻度が低いものより高いものであるという予想から，同音異義語辞書内を出現頻度でランキングし上位 5 件からランダムに選択する．加えて，文中の修正箇所は 3 箇所までとし，この数を超えないようにランダ

[1] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

試験区分: 情報科学区分
氏名: 夏見 昂樹
希望研究室: 自然言語処理学研究室
現在の専門: データマイニング

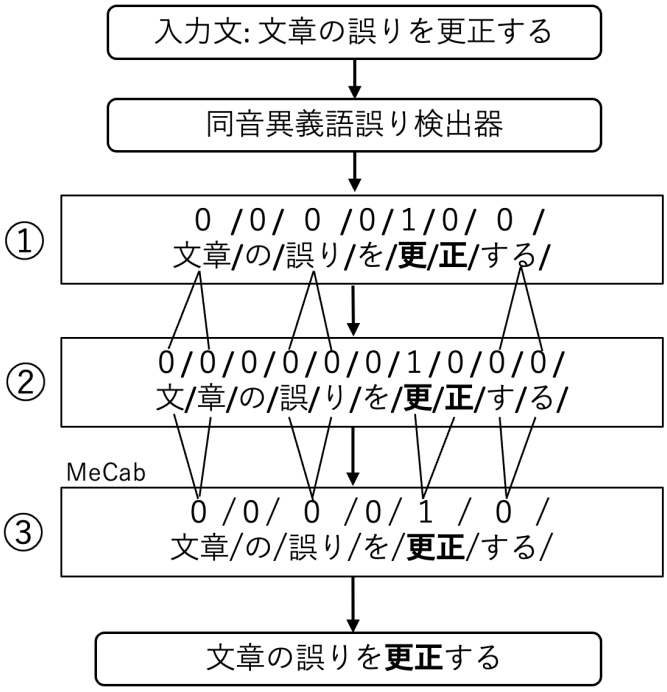

図 1: 同音異義語誤り検出手法の概略図

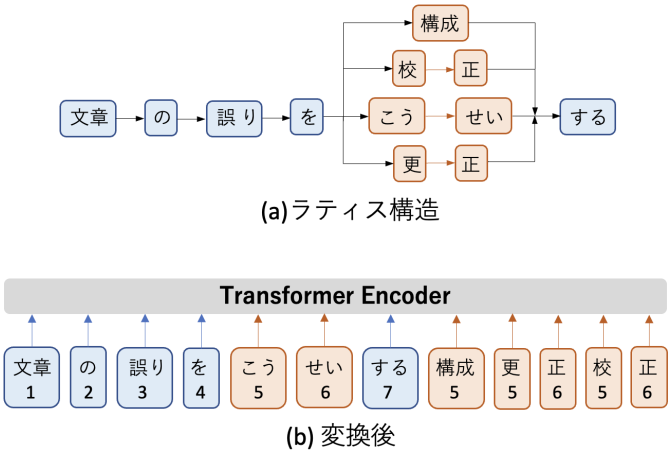

図 2: ラティス構造と変換後の入力シーケンス

ムに選択する[2]．作成された正文と誤り文を用いて各トークンごとに 0(誤りなし) と 1(誤りあり) の二値分類教師あり学習を BERT に対して行う．

図 1 は同音異義語誤りの検出手順の概略を示しており，作成した検出器からは ① のような出力が得らる．これを MeCab の分割に当てはめることで同音異義語誤り部を特定したいが，出力はサブワードごとになるため MeCab の分割とサブワードの分割が一致するとは限らない．そこで ② のように一度文字レベルに 0,1 の割り振りを行う．最後に ③ で MeCab の分割と検出器の出力を照らし合わせ，各分割ごとに論理和の処理を行うことで同音異義語誤り部の検出が可能となる．

[2]これらのパラメータについては検討する必要がある

2.3.2 同音異義語誤り部へのラティス構造の付与

検出された同音異義語誤り部に対して，ラティス構造を適応したシーケンスの作成手順について述べる．まず，藤井ら [4] の手法に基づき，誤り部を BERT の [MASK] トークンに置き換えて，BERT に [MASK] 部を予測させる処理を行う．予測結果を BERT の尤度順に 10,000 語取得し，それらが [MASK] 部分の読みと対応する同音異義語辞書中に存在すれば，上位から最大 3 件抽出する[3]．図 2(a) では上記の処理によって得られた同音異義語候補を含んだラティス構造の例を示している．この際，Qin ら [2] の入力シーケンス中に音節情報を組み込む手法を参考に，提案手法では日本語の読みをラティス構造に組み込む．これによって抽出した同音異義語候補内に正例がなかった場合であっても，NMT モデルが日本語の読みに注視して学習することを期待する．

2.3.3 Transformer へのラティス構造の入力・訓練

ラティス構造を NMT モデルへの入力を可能とするために，図 2(b) のように，ラティス構造から平文に変換することで NMT モデルへの入力を可能とする．また，NMT モデルへのサブワード入力を可能とするために，同音異義語候補内の，最長サブワード分割数の候補を基準に，左詰めで他の候補の位置情報を振り分ける．これは同音異義語候補ごとにサブワード分割数が一致するとは限らないためである．

JESC，JPO などの対訳コーパスの日本語入力文に対して，提案した検出器を適応し，NMT モデルを訓練する．

2.4 提案手法の評価

学習した NMT モデルの翻訳性能を BLEU などによって測定し，従来の Transformer との翻訳性能を比較することで，提案モデルの翻訳性能を評価する．

提案した NMT モデルが，同音異義語誤りに対して効果的か評価する必要があるが，実世界での日本語の同音異義語誤りのみを取り扱った対訳データは存在しない．そこで，Qin ら [2] に基づいて，テストセット内に一定の確率で人工的に同音異義語ノイズを混入させ，翻訳性能を評価することで NMT モデルの同音異義語誤りへの頑健性の評価を行う．

参考文献

[1] Y. Belinkov et. al., ”Synthetic and natural noise both break neural machine translation”, ICLR, 2018.
[2] W. Qin et. al., ”Modeling Homophone Noise for Robust Neural Machine Translation”, CASSP, 2021.
[3] 三品 et. al., ”確率的 LSA を用いた日本語の同音異義語誤りの検出・訂正”, 情報処理学会論文誌, 2004.
[4] 藤井 et. al., ”BERT を利用した Zero-shot 学習による同音異義語の誤り検出”, 言語処理学会 第 27 年次大会, 2021.
[5] A. Vaswani et. al., ”Attention is all you need”, NeurIPS, 2017.
[6] F. Xiao et. al., ”Lattice-Based Transformer Encoder for Neural Machine Translation”, ACL, 2019.
[7] X. Li et. al., ”FLAT: Chinese NER Using Flat-Lattice Transformer”, ACL, 2020.

[3]本稿では同音異義語候補を抽出する際に BERT の MLM を用いたが，純粋に同音異義語の出現頻度が高いものを数件用いるなど他の手法についても実験・評価する必要がある．