ディープラーニングを用いた誤字・脱字検知システムの研究

東京情報大学　総合情報学部　総合情報学科
知能情報システム研究室　永井ゼミ4年
J20387　前原 龍平

令和5年度　卒業研究 II

目次

- 研究の動機と目的
- 既存の関連研究
- 研究環境
- システムイメージ、開発手順
- システムの利用例
- 研究計画
- 現在の進捗(2023/06/21)

研究の動機と目的

- SNSの投稿やメールの送信をする際に、誤変換などによる誤字・脱字があるまま送信してしまうことがある

→防ぐ、または頻度を減らすことができないか？

既存の関連研究(下線部リンク)

電子ジャーナルプラットフォーム
「科学技術情報発信・流通総合システム(J-STAGE）」より引用
閲覧日時：2022/11/16

- **Bidirectional LSTMを用いた誤字脱字検出システム**
  高橋 諒, 蓑田 和麻, 舛田 明寛, 石川 信行
  株式会社リクルートテクノロジーズ, 株式会社PE-BANK
  雑誌の原稿から学習データセットを作成し、BLSTMモデルを用いて「誤字脱字箇所」「正しい文字の候補」「誤字脱字を含む文」を出力する

既存の関連研究(下線部リンク)

電子ジャーナルプラットフォーム
「科学技術情報発信・流通総合システム(J-STAGE）」より引用
閲覧日時：2022/11/16

- **日本語 Wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの構築**
  田中 佑, 村脇 有吾, 河原 大輔, 黒橋 禎夫
  Wikipediaの編集履歴から文字単位で入力誤りを抽出しデータセットとして使用
  seq2seq(sequence to sequence)モデルを用いて、入力誤りと漢字の読み推定を同時に学習

既存の関連研究(下線部リンク)

電子ジャーナルプラットフォーム
「科学技術情報発信・流通総合システム(J-STAGE)」より引用
閲覧日時：2022/11/16

- **ベイズ理論を用いたキー入力の間違い推定法の検討**
  西村 希槻, 寺澤 卓也
  キーボード入力から入力ミスを検知する

既存の関連研究(下線部リンク)

電子ジャーナルプラットフォーム
「科学技術情報発信・流通総合システム(J-STAGE)」より引用
閲覧日時：2022/11/16

- **自然言語処理手法に基づく作文自動点検の試み：学習指導要領「書くこと」に立脚した文章評価補助のための技術(自由研究発表)**
  藤田 彬
  コンピュータからの入力を基に文法ミスを検出

研究環境

- **Jupyter Notebook**
  無償で利用できるディープラーニングの学習環境
  Google Colaboratoryでも使用可
  プログラミング言語はPython
  Webブラウザで動作、GPUも使用可能

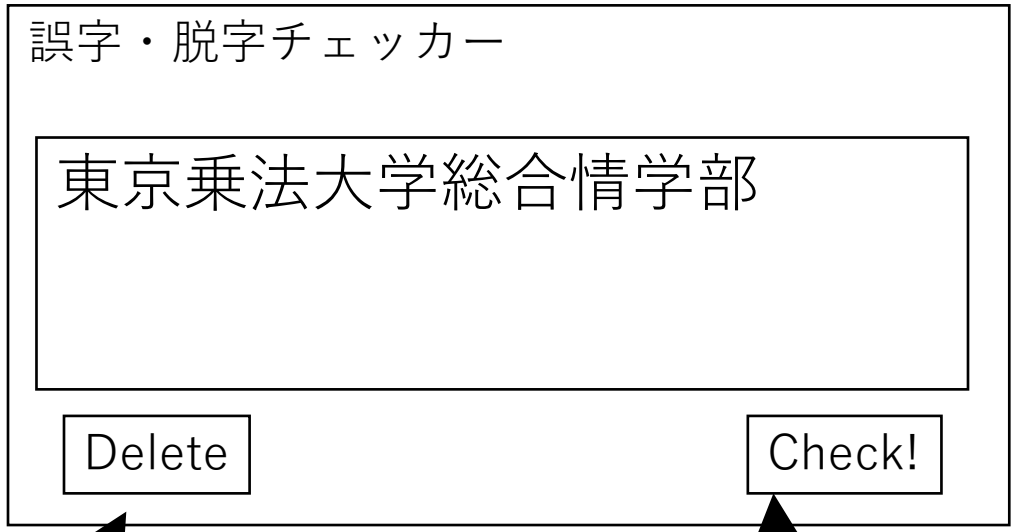
システムイメージ

- ~~**RNN(Recurrent Neural Network)**モデル~~　**予定変更**
  使用モデルの選定中、最有力候補は**BERT**
  ※Googleが開発した自然言語処理のアルゴリズム
  長く複雑な文章を理解でき、Google検索に利用されている
  より軽量で高速な**ALBERT**も存在するが、そちらはどうか？

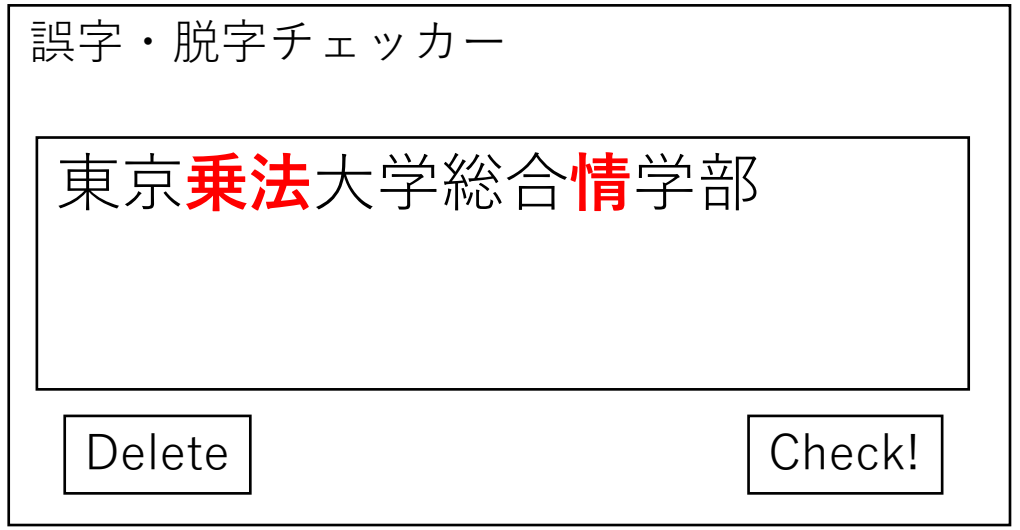
- 特定の単語が入力されている場合に、
  その前後に入力される可能性の高い単語・低い単語を予測
  入力される可能性の低い単語が入力された場合、
  誤字・脱字として検出する

システムイメージ(インタフェース)

①テキストエリアに入力



③誤字・脱字を強調表示



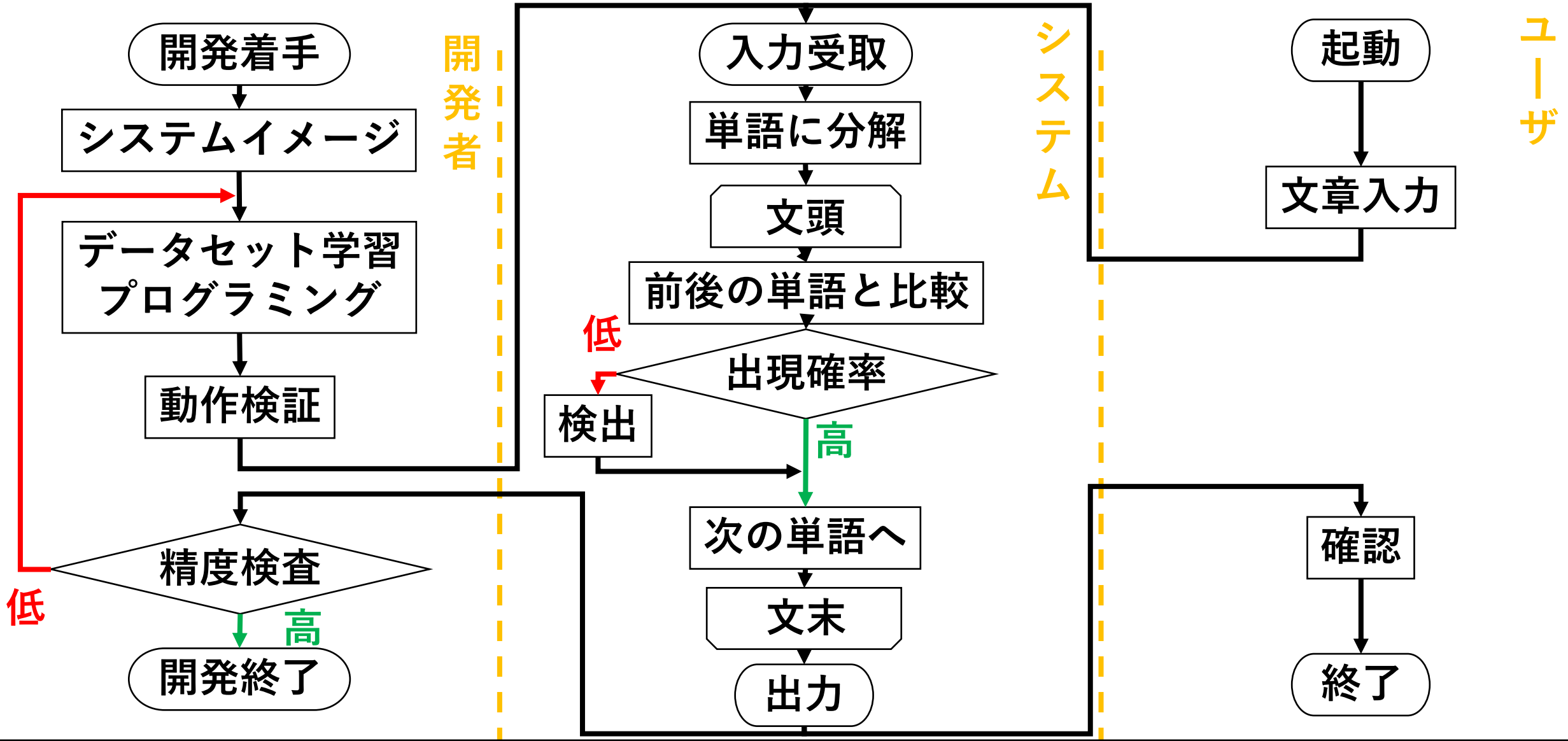
②「Check!」ボタンを押す

「Delete」ボタンを押すと、
テキストエリア内の文章を削除

変更案→
- 下線？マーカー？
- 正しい案の表示？
- 自動修正？

開発難度 高

システムイメージ


開発着手
システムイメージ
データセット学習
プログラミング
動作検証
精度検査
低
高
開発終了
開発者
入力受取
単語に分解
文頭
前後の単語と比較
低
出現確率
検出
高
次の単語へ
文末
出力
システム
起動
文章入力
確認
終了
ユーザ

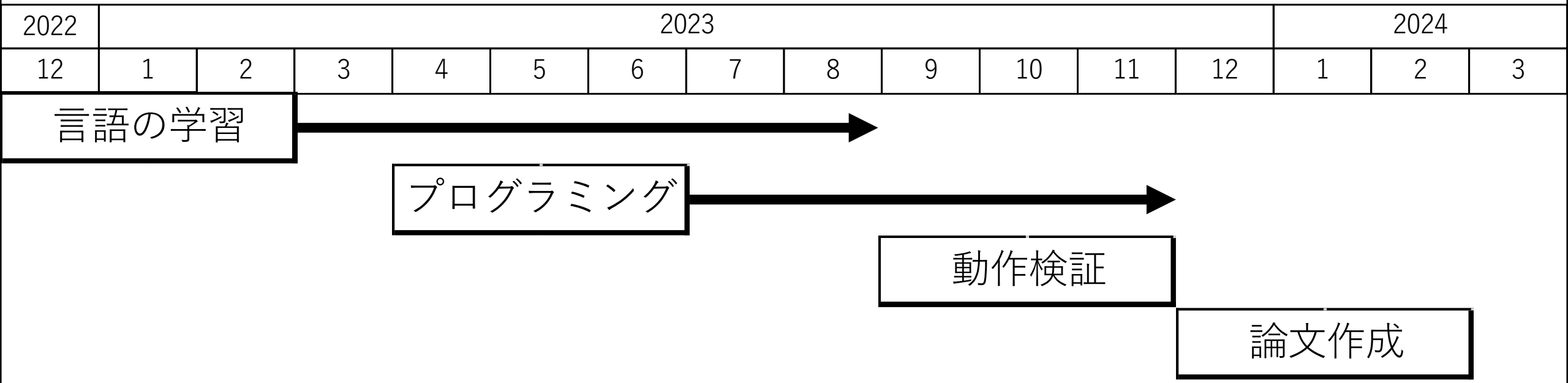
システムの利用例

- **メール、SNS、掲示板サイトなど**
  特に、一度書き込んだら取り消せないものは、入力ミスが残り続ける
  そうでなくても、気づく→削除→再度送信、という手間が……

- **レポートや自己PR文** (学生)
  誤字・脱字が減点対象となることもある

- **書類の作成** (事務職)
  誤字脱字による情報の伝達ミスのリスク

研究計画


2022
2023
2024
12
1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
言語の学習
プログラミング
動作検証
論文作成

現在の進捗・課題

①文章を入力するインタフェースを作成　**成功**

↓

②文章を単語に分解するプログラムを作成　**成功**
　Janome(Pythonの形態素解析エンジン)を使用

↓

③単語ごとに、隣接する単語との繋がりが正しいかどうかを判断するプログラムを作成
　今後の課題

↓

④結果を出力するインタフェースを作成
　②の結果の出力は成功

現在の進捗・課題

- 今後やるべきこと
  使用するモデル・データセットの選定
  (現時点での使用予定モデルは**BERT**または**ALBERT**、
  データセットは**Wikipediaデータベース**)
  誤字・脱字を検知するプログラムの作成
  検知した単語を強調表示するプログラムの作成

- 今後の予定
  プログラミング、および動作の検証：年末まで