タイポ修正を含む頑健な形態素解析

B4　丸谷　怜史

12/5

タイポ修正

入力文の打ち間違いを訂正し出力

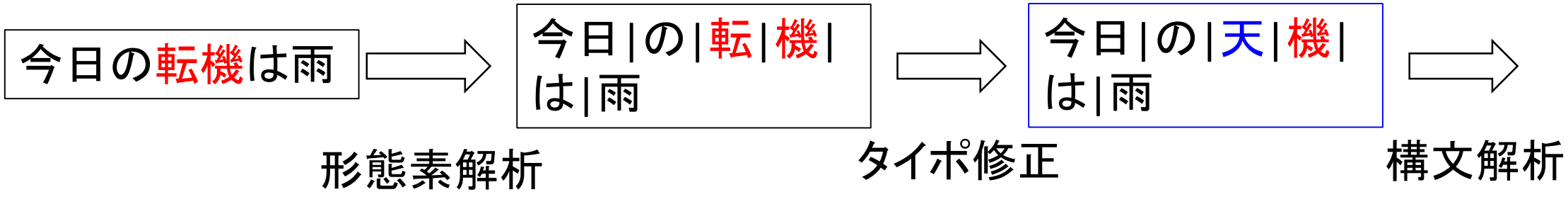
ex)今日は素晴らしい転機だ。
→今日は素晴らしい天気だ。

脱字、衍字にも対応

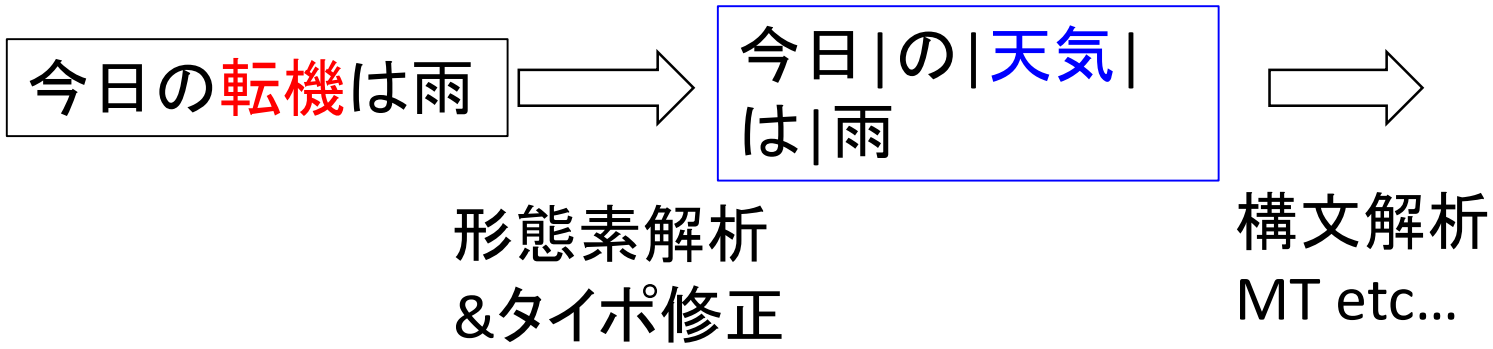
ex)日頃ろの行いを顧よ。
→日頃の行いを顧みよ。

提案手法

従来


今日の転機は雨
今日|の|転|機|
は|雨
今日|の|天|機|
は|雨
形態素解析
タイポ修正
構文解析
MT etc…

提案


今日の転機は雨
今日|の|天気|
は|雨
形態素解析
&タイポ修正
構文解析
MT etc…

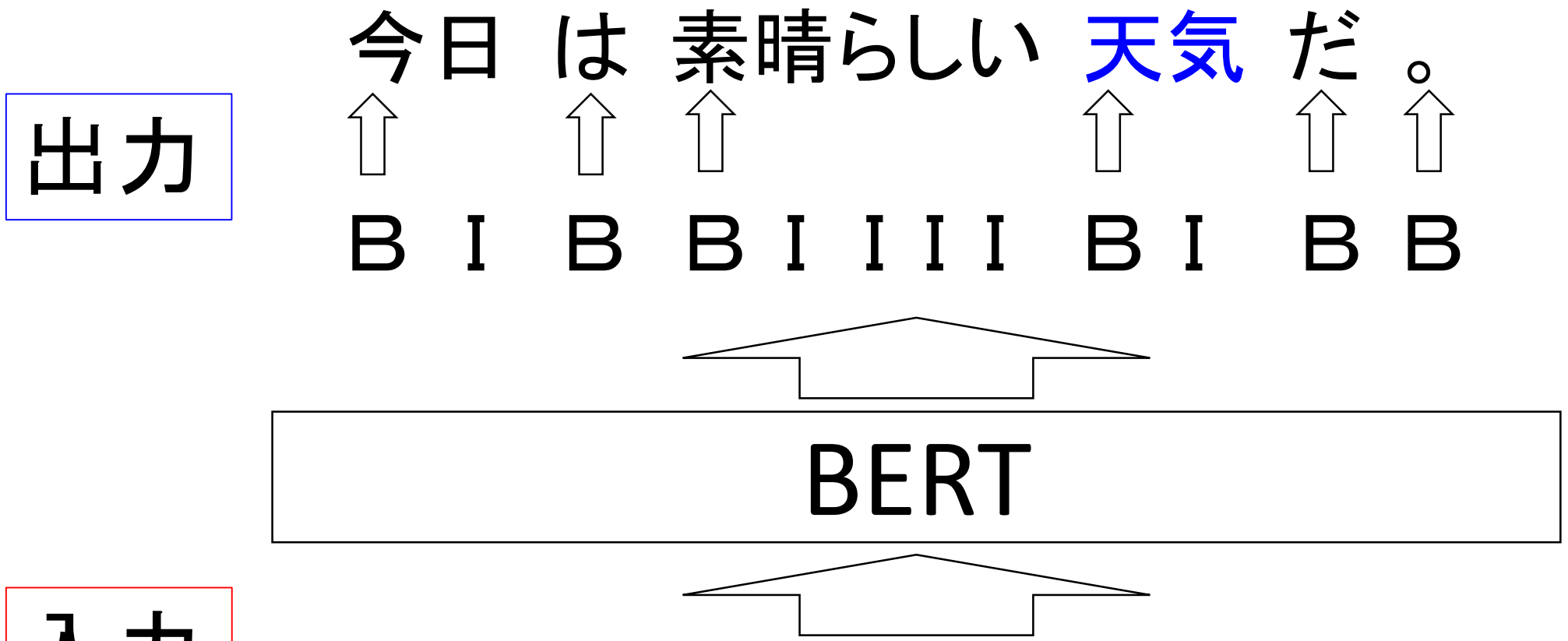
関連研究

如何にしてデータの無駄を省き、速く学習させるか

・Shrinking Japanese Morphological Analyzers With Neural Networks and Semi-supervised Learning[Arseny+ 2019]

・Encode, Tag, Realize: High-Precision Text Editing[E.Malmi+ 2019]

BERTに基づく単語分割とタイポ修正の統合モデル



単語分割の実験設定

- コーパス：KWDLC
  - train.txt 381,488文字、12,578文
  - dev.txt 43,339文字、1,388文
  - test.txt 36,988文字、1,175文
- 日本語文字単位BERT Pretrainedモデルを使用
- エポック数：3

単語分割:結果

モデル	F値
JUMAN	98.1
JUMAN++	98.6
Arseny+	98.7
提案手法	97.6

単語分割：誤り分析

・「形容動詞」と「名詞＋助動詞」

ex)出力「だめ|です|ね」　　正解「だめです|ね」

「早々に」　　「早々|に」

・漢字が主の単語内に平仮名が入っているもの

ex) 出力「日|の|出」　　正解「日の出」

単語分割：誤り分析

・片仮名の単語の境目

ex)出力「ラフス|ライス」　正解「ラフ|スライス」

・人間が判断しても悩ましいもの

ex) 出力「とらふぐ」　　正解「とら|ふぐ」

今後の課題

- タイポ修正モデルを構築し実験
  - コーパスは準備済み