大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

BERTを用いたトラブル調査記事のラベリング

富士通株式会社　　白木 康建, 野呂 惇

開発における問題点

・企業活動においてトラブル対応は重要
　・予測困難かつ多発するため，トラブル対応による生産性の低下が問題
　→トラブル発生を抑止することが求められている
・発生頻度の高い既知のトラブルへの対策が効果的
　・現状：トラブル調査記事(*1) を元に1件ずつタグ付け/分類することで頻度の高いトラブルを選定
　・問題点：精度が低い/コストが高い

*1: トラブル対応の履歴

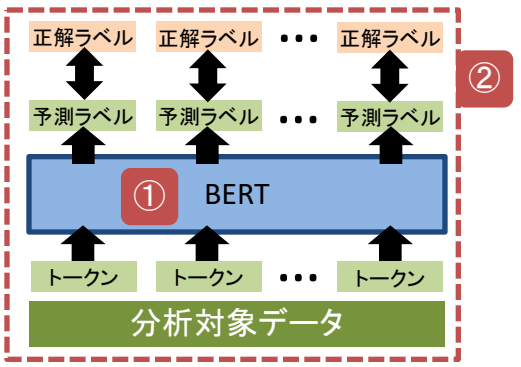
手法・ツールの適用による解決

・固有表現抽出技術を使用し，トラブル調査記事から，「現象/疑問点」，「原因」，「対処方法」の情報を抽出
　・固有表現抽出による機械的な抽出で期待する効果
　　・精度の向上
　　・コスト削減
・固有表現抽出：
　テキストから固有表現（人名、組織名等）を抽出するタスク（=ラベリング）

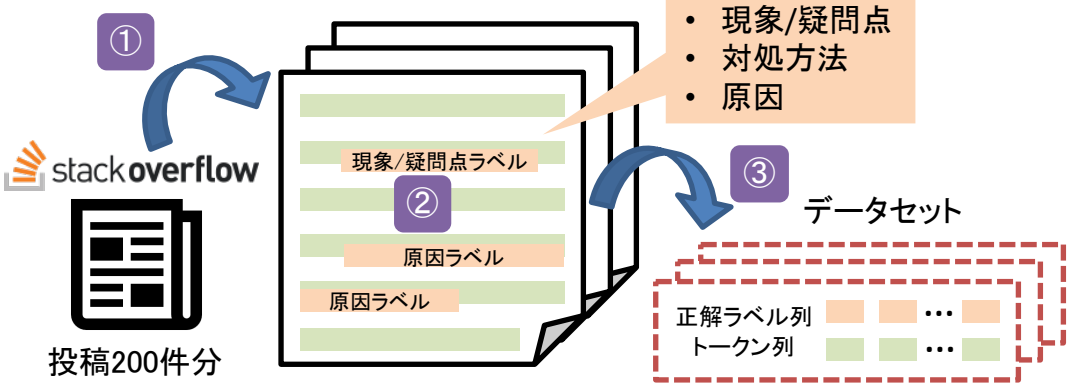
固有表現抽出タスクの実装

実装方法
①公開されている事前学習済みBERTモデルを活用
②固有表現抽出タスク用にファインチューニング

BERT
- Google社が開発した自然減処理モデル
- 各種自然言語処理タスクへの高い汎用性

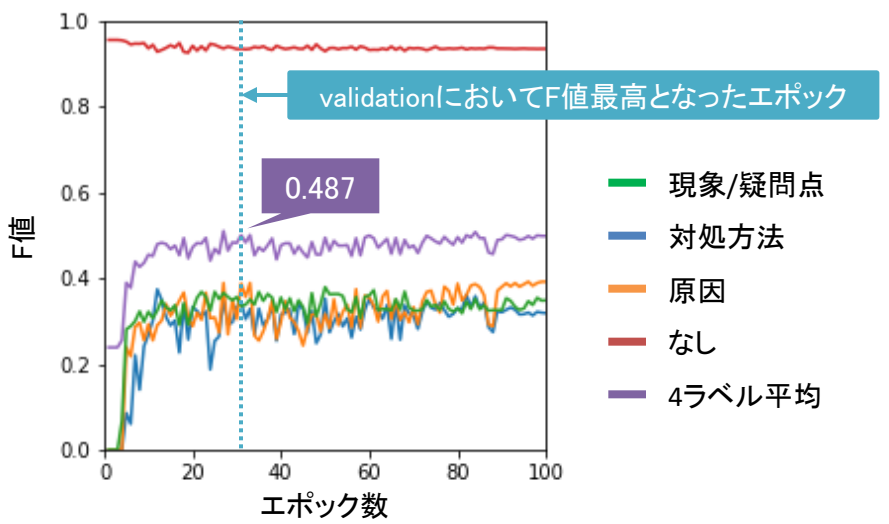


データセット作成手順
①日本語stack overflowの投稿から分析用テキストを抽出
②手作業でテキストにラベリング
③テキストをトークンに分割

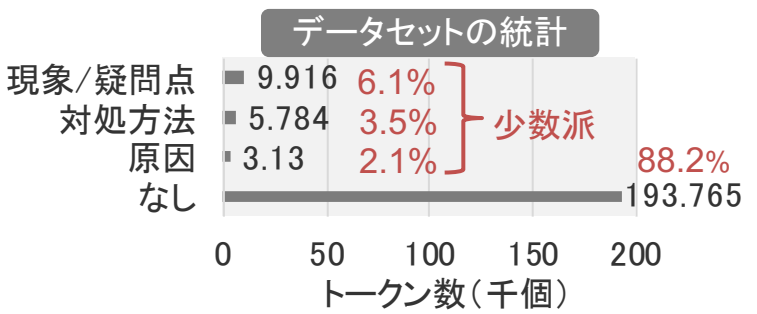


実験結果

トークン毎のラベル予測結果をF値にて評価



- 目標（F値 > 0.8）未達だが学習が行われていることを確認
- 少数派ラベルの検出精度が低く不均衡データへの対策が課題



考察

左記の通り，期待する精度を得ることはできなかった．
学習に使用するデータ数不足，学習のアルゴリズムやパラメータ調整が不十分であったことが原因と考える．
今後の課題は以下．
・ラベリング精度に関する課題
　・オーバーサンプリングの実施，構造情報や文脈を加味した学習，データセットの追加
・評価方法に関する課題
　・クロスバリデーションによる汎化性能評価が必要．