SecHack365

テキスト推論認識を用いたファクトチェッカー「RTE Fact Checker」

思索駆動コース　西尾勇輝

概要

社会に蔓延する誤情報・偽情報への対策として，WEB上の文章の各文が信頼できる文章と矛盾がないかどうかをRTEと呼ばれる手法を用いて判定するシステム，RTE Fact Checkerを作成した．信頼できる文章，判定する文が共に数文程度であれば，真偽の判定結果を現実的な時間で返すことができる．

偽情報・誤情報への対策

人手によるファクトチェック

• 一定の基準に基づいて検証，発表しているため信頼性が高い． • オフラインの情報源からも情報を収集できる．	• 社会的に重要な言説に限られる． • 疑わしい情報と検証・発表の間に時間差が生じる．

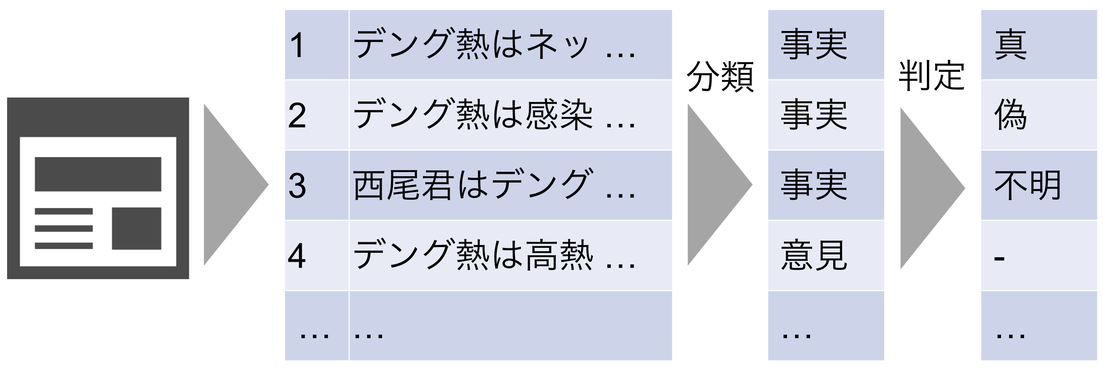
既存のフェイクニュース検出ソフトウェア

• 低コスト • リアルタイムで判定できる．	• 記事全体で判定しているため，どの言説が正しくないのか分からない．

個別の事実をファクトチェックするシステムが必要！

RTE Fact Checker

個別の事実をファクトチェックするために，記事の各文が信頼できる情報源の文章と矛盾していないかどうかを判定するシステム．対象の文章を文に分割し，意見を述べている文と事実を述べている文に分類する．信頼できる文章に基づいて，事実を述べている文の真偽判定を行う．



意見と事実の分類

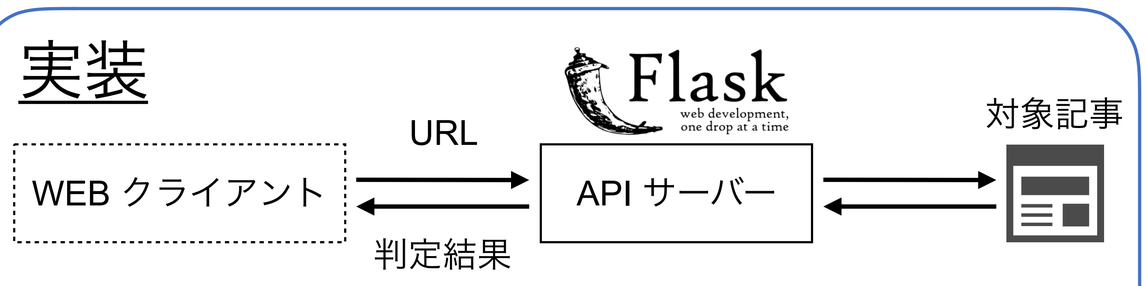
松林(2019年度研究駆動コーストレーニー)の機械学習モデルを使用．文を形態素に分け，文中の品詞の割合等を用いて意見を述べている文と事実を述べている文に分類する．351文を分類し，学習データを作成した．

RTEによる真偽判定

RTE(テキスト推論認識)をするシステムccg2lambda[1]を使用．ccg2lambdaは前提文から仮説文が推論できるかどうかを判定する．推論できれば”yes”，矛盾すれば“no”，判定ができなければ“unknown”を返す．前提文は複数でも可能．

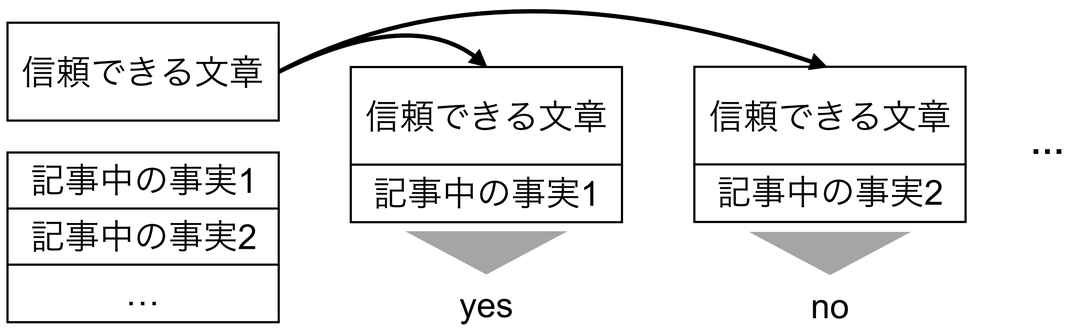
例　前提文: ヨーロッパ出身の参加者はいない．
　　仮説文: ヨーロッパ出身の女性参加者はいない．
　　→ yes

実装



Flaskを用いてRESTful APIサーバーを実装した．APIサーバーは，受け取ったURLの記事を解析し，判定結果を返す．

信頼できる情報源からの文章を保持しておき，それらの文章をRTEにおける前提文にする．事実に分類された各文を仮説文とし，推論を行う．



例

4文からなるWEBサイトを作成し，真偽判定を行った．信頼できる情報源として厚生労働省検疫所のサイト[2]の4文を使用した．

信頼できる文章

- デング熱はデングウイルスによる感染症でネッタイシマカやヒトスジシマカによって感染します。
- 感染症法の4類に分類されています。
- ウイルスを持っているネッタイシマカやヒトスジシマカなどに刺されることで感染します。
- ヒトスジシマカは、ヤブ蚊とも呼ばれ、日本にも生息しています。

結果

レスポンス速度: 74.94 (秒)

	記事中の文	分類	判定
1	デング熱はネッタイシマカによって感染する。	事実	yes
2	デング熱は感染しない。	事実	no
3	西尾君はデング熱で入院した。	事実	unknown
4	デング熱は高熱が出てとても辛かった。	意見	-

展望

- ccg2lambdaに医療や金融といった特定の分野の知識及び新たな推論規則を追加することで，RTEの精度を上げていく．
- RTE Fact Checkerを用いて，ある分野の情報をまとめたWEBサイトを分析する．
- 信頼できる情報源をユーザーが追加できるようにすることで，用途に合わせて使いやすいシステムにする．

[1] Pascual Martínez-Gómez, 2016
[2] 厚生労働省検疫所，FORTH 海外で健康に過ごすために(https://www.forth.go.jp/useful/infectious/name/name33.html)

NICT 国立研究開発法人 情報通信研究機構