研究紹介

ことばをコンピュータで処理する技術とその応用

香川大学　創造工学部　創造工学科
情報システム・セキュリティコース

安藤研究室

当研究室の研究分野 と 扱うデータ

- **研究分野を一言でいえば**
  - **人間のことば（＝自然言語）をコンピュータで処理したり，理解する技術と，その技術を応用する研究（自然言語処理）**
    - つまり，**人工知能を作る研究**
- **自然言語に関するデータ**
  - **話す　→　音声データ**
  - **読む，書く　→　テキストデータ　※こっちが研究対象**
  - 話したことば（音声データ）をコンピュータが認識する技術（音声認識）を使えば，**テキストデータに変換可能**
  - **つまり，テキストデータは，ほぼ無限に存在！**

当研究室の研究目的

- **自然言語は**
  人間社会における**代表的なコミュニケーション手段**
- **人間社会において**
  テキストデータは，今この瞬間も増え続けている
  - 医療・福祉，教育，ビジネス，エンタメなど，
    人間社会のどの分野をみてもテキストデータが必ず存在
  - **よって，人間社会には，**
    自然言語に絡む様々な社会問題やニーズも多数存在
- **当研究室の研究目的は**
  社会問題の解決や社会のニーズに応える
  自然言語処理の要素技術やシステムの実現！

近年の当研究室の研究例（抜粋）

1. **SNS上の違法薬物売買アカウントの検出**
   （**犯罪者を見つけ出せ！**）
2. **ツイートされる病気・症状の可視化**
   - 病気・症状の事実性判定（**本当に病気・症状？**）
   - 投稿者の居住地・活動地域推定 （**どこに住んでいる？**）
3. **小学校教師に対するNIE支援**
   - 新聞記事の内容を補足する画像推薦（**難しいなら補足！**）
   - NIEに適した新聞記事の推薦 （**適切な記事を推薦！**）
4. **現地でしか買えない土産の情報収集と提示**
   （**レアな土産の選択を支援！**）
5. **ファンタジー小説からの登場人物情報の抽出**
   （**こんな小説が読みたい，人物関係が知りたい等を解決！**）

※他にも色々取り組んでいる

1. SNS上の違法薬物売買アカウントの検出

背景と目的

- 若者が普段使用しているSNS上で薬物売買が盛んになり，違法薬物に手を出す若者が増加
- **違法薬物売買アカウントを自動検知することで，違法薬物の売買防止を実現**

売買ツイートの特徴例

- 取り締まりから逃れるため，「**野菜**」や「**アイス**」などの**日常利用する単語の隠語**や**絵文字**などが利用

提案手法の概要

- 機械学習を用いて，**売買投稿から投稿内容の特徴や絵文字タイプなどを学習**することで違法薬物売買を判定
  - 現在，提案手法により，投稿内容から**約98%の性能**で違法売買に関する投稿を判定可能

2. ツイートされる病気・症状の可視化 - 病気の事実性判定 -

背景と目的

- Twitter上では，様々な病気・症状がツイートされている
- **いつ，どこで，どのような病気・症状がツイートされているのか，何が起こっているのかを，地域別・時系列別に可視化するシステムの構築**
  - **感染症の流行把握だけでなく，未知の病気の発生も検知可能**
- **実現には，様々な病気・症状の事実性を判定する技術が必要**
  - 例1) 明日，試験で頭が痛い：**偽**，例2) 朝から熱がでて，頭が痛い：**真**

提案手法の概要

① 病気症状ツイートの真/偽事例を収集し，**ツイート内の単語や文末表現，主体などの特徴を学習**した判定器を構築

② 判定器で，病気・症状を含む新しいツイートの真偽を判定
  - 現在，提案手法により，**約77%の性能**で病気症状の事実性を判定可能

2. ツイートされる病気・症状の可視化 - 投稿者の居住地・活動地域推定 -

背景と目的

- Twitter上では，様々な病気・症状がツイートされている
- **いつ，どこで，どのような病気・症状がツイートされているのか，何が起こっているのかを，地域別・時系列別に可視化するシステムの構築**
- **地域別の可視化を実現するには，ツイート投稿者の居住地や活動エリアなどを推定する技術が必要**

提案手法の概要

① 居住地等を明示しているユーザのツイート集合を収集し，**地名や活動情報などの特徴を学習**した推定器を構築

② 病気・症状ツイートを検知し，そのユーザのツイートを収集

③ 推定器を用いて，病気症状ツイート投稿者の居住地を推定
- 現在，提案手法により，**約73%の性能**で47都道県別の居住地を推定可能

3. 小学校教師に対するNIE支援 - 新聞記事の内容を補足する画像推薦 -

背景と目的

- 小学校では，新聞を活用する教育（NIE）が実践
- しかし，小学生にとって新聞記事はハードルが高い
  - **理由**：内容が難しい，単語・表現が難しい，**内容理解を補足するための図や写真などもほとんど付与されていない等**
  - **教師は，通常業務に加えて，補助教材の準備が必要となり，負担も増加**
- **教師が選択したニュース記事に対し，記事の内容を補足する画像コンテンツを推薦するシステムを構築**

提案システムの概要

① 入力記事から検索語を自動生成し，Web上から画像検索

② 画像タイプや記事内容との関連性，解説性などでスコアを計算し，そのスコアを基に補足画像をランキングして提示

3. 小学校教師に対するNIE支援
- NIEに適した新聞記事の推薦 -

背景と目的

- 小学校では新聞を活用する教育（NIE）が実践
- **5科目を担当する小学校教師にとって，NIEに適した記事の選択コストは非常に大きい**
- **日々発行され続ける膨大な記事の中から，NIEに適した記事を教師に推薦するシステムを構築**

提案システムの概要

① NIEで実際に利用された or 専門家が選択した記事群を収集
- このような記事には**NIEに適した何かしらの特徴が含まれている**と仮定

② 収集した記事群から**記事タイプ，読みやすさ，内容などの記事特徴を学習**し，新しい記事に対し，NIEでの適切性を判定
- 現在，提案手法により，**約94%の性能**でNIEに適した記事を判定可能

4. 現地でしか買えない土産の情報収集と提示

背景と目的

- 様々な土産がネットショッピングで購入可能になり，その一方で，**現地でしか買えない土産の価値が高まっている**
- **現在，現地でしか買えない土産に関する情報を提供するサービスやアプリは存在していない**
- **そこで，Web上に散在する「現地でしか買えない土産」に関する情報を収集し，提示するシステムの構築を実現**

提案システムの概要

① Blog等から収集した土産に関するテキストに土産情報をタグ付けし，その**テキストから特徴を学習**した抽出器を構築

② Blog等から土産に関する新しいテキストを収集し，抽出器を用いて，**土産名，店舗名，評判などの情報を抽出**

③ 対象の土産が現地でしか買えないのか，手に入りにくいのか等の観点でスコアリングし，土産情報と共に提示

5. ファンタジー小説からの登場人物情報の抽出

背景と目的

- **小説を探す，読むときの問題**として
  - ”大学教授が異世界に転生し，魔法使いとして活躍する小説が読みたい“
  - 隙間時間で読んでるから，人物関係がよくわからなくなった

  など，**登場人物に関係するものが存在**
- 問題の解決には登場人物情報を抽出し，活用する技術が必要
- **ファンタジー小説から登場人物情報を抽出する技術を実現**
  - 登場人物情報：人物名，性別，年齢，所属，地位，能力など

提案手法の概要

① 登場人物情報をタグ付けした小説テキストから，**各種の特徴を学習した抽出器**を構築

②抽出器を用いて，新しい小説テキストから人物情報を抽出

- 提案手法により，現状，**約88%の性能で人物名**を，**約80%の性能**で**地位**などを抽出可能

最後に

- **自然言語は，我々に生活に必要不可欠であるからこそ，解決すべき課題も多い**
  - **我々の生活に直結した「役に立つ」研究に取り組める**
- **提案する基礎・要素技術が，人間が自然言語を理解する仕組みの解明にもつながる**
  - **人工知能の実現**
- **自然言語は手強い研究対象であるが，必要不可欠で非常に興味深い研究対象でもある**
- **是非，一緒に研究しましょう！**