自然言語処理
第1回

中田尚

スケジュール

- 火曜日3限
  - 4/8から7/29
  - GWで2週連続休み（4/29と5/6）

授業の到達目標及びテーマ

- 人間の言葉である自然言語を計算機処理の対象として、電子化された膨大な情報資源から有用な知識を取り出すテキスト情報検索や、異なる言語間のコミュニケーションを支援する機械翻訳などの技術の基礎となる概念や技法について、理解を深め実践技法を習得する。
- 日本語や英語などの自然言語を対象として計算機処理を行う上で前提となる、言語固有の構造と様々な処理レベルの基本概念を理解するとともに、情報技術の専門職として心得ておくべき各種技法について実例を通して学ぶ。具体的には、次の事項を到達目標とする。
  - 自然言語を取り扱う上で基本となる言語学的な種々の概念を簡潔に説明できる。
  - 各種のツールを用いた形態素・統語・意味解析手法を理解し、例示できる。
  - 自然言語処理の応用先である、情報検索やテキストマイニング、機械翻訳等の仕組みを理解し、公開されているソフトを用いた処理が行えるようになる。
  - 教員の指導の下に自然言語処理に関する課題を設定し、その解決手段を考案する。

テキストと評価

- テキスト
  - 『自然言語処理の基礎』、奥村学著、コロナ社、ISBN：978-4-339-02451-7
  - 本日配布
- 学生に対する評価
  - 科目認定条件
    - ※出席率について80％以上であること。
    - ※定められた提出物が80％以上提出されていること。
  - 科目評価方法
    - 小テスト20%・グループワークの貢献度20%及び期末レポート60%
    - 小テスト及び期末課題レポートなどを総合して評価。

ChatGPTのレポートでの利用（去年と同じ）

- 参考にして引用とするのであれば全く問題なし
  - ChatGPTにXXXと質問したらYYYという答えだったので、私はZZZと考えました。
    - ZZZがなければ0点
- YYYの部分だけのコピペ
  - なんとなくわかる場合はあるが、正直わからない
  - 疑問を持たれて評価が低くなるかもしれない
- 結局は自分の勉強になっていれば良い
- **注：他の授業については各先生の指示に従うこと**

生成AIとの付き合い方

- 生成AIを使わないのはもったいない
  - 大学生の50%が利用（2024年末）
    - 2023年末は30%
  - 自転車があるのに走るようなもの
- 大学の課題は生成AIで解ける
  - 東大入学の能力があると言われている

- 東大生の家庭教師がいるようなもの
  - 課題の答えも全部教えてくれる
  - 家庭教師に宿題を解いて貰う？？

ChatGPTの利用例

【問題文】
高さ h = 1.0 m の位置から静止状態でピンポン玉を自由落下させる。
落下中に高さ H の位置に板が設置されており、板は地面に対して角度 θ で傾いている
(0 < θ < π/2)。

- ピンポン玉は板に衝突し、反射係数 e = 0.9 に従って跳ね返るものとする。
- 空気抵抗は無視できるものとする。

ピンポン玉が原点から地面に到達するまでの水平方向の距離 x を一般式で求めよ

DISCO
DECC
DISCO EQUIPMENT
CODING CONTEST

ChatGPTの解答

まず、ピンポン玉が高さ h から落下する際の衝突直前の速度 v_{before} を求めます。
自由落下の式より

$$v_{\text{before}} = \sqrt{2gh}$$

（途中省略）

$$x = \left(v_{\text{before}} \cos^2\theta + e v_{\text{before}} \sin^2\theta\right) \cdot \frac{(1-e)v_{\text{before}} \sin\theta\cos\theta + \sqrt{(1-e)^2 v_{\text{before}}^2 \sin^2\theta\cos^2\theta + 2gH}}{g}$$

これが水平方向の移動距離 x の一般式です。

生成AIとの付き合い方

- すでに大学生レベルのことは生成AIで可能
  - やや誇張ではあるが、2年後には誇張でない可能性は高い

- 現在発生している問題
  - 生成AIの出力が理解できない
    - 正しいか判断できない
    - 実行してみたらちょっと違うんだけど、修正できない
    - 答えが書いてあるのにそれが答えだということに気付かない
  - 使う人のスキルでできる/できないが変わる
    - A：「生成AIにやって貰おうと思ったんだけど〇〇ができなかった」
    - B：「私がこう指示したらできたけど」
    - A：「・・・（やり方見ても理解できない）」

自然言語処理

自然言語とは（↔人工言語）

- 人間が日常的に使用している/いた言語
  - 文法よりも先に言語が存在している
  - 地域/時代によって意味や文法が変化することもある

- ラテン語
  - ローマ帝国の公用語だが現代では日常会話に使う人はいない
  - 学術的には使われている（植物の学名など）

自然言語における例外の例

- 英語の動詞の不規則変化
  - eat-ate-eaten
- 英語の複数形
  - man/men、datum/data
- 発音の例外
  - ついたち、ふたり

- 理由があるとしても基本的に後付け

例外を排除した人工的な自然言語

- エスペラント語
  - 「エスペラントは1880年代にルドヴィコ・ザメンホフによって創案された。最初の文法書・単語集は1887年に発表された。」（Wikipedia）
  - 世界の言語を統一しようという考え方

人工言語

- 国際補助語
  - エスペラント語
- 形式言語
  - プログラミング言語
- 架空言語
  - ハンター語（HUNTER×HUNTER）
- この授業の範囲外だが対比として使うことはある

自然言語処理

- 自然言語を機械的に処理する記述
  - コンピュータにことばを理解させる
  - コンピュータ上でことばを処理する

- 文字、音声、映像（手話）を含むが、ここでは主に文字で記述された文を対象とする

自然言語処理の応用

- 機械翻訳
- 対話、インターフェース
  - ELIZA（イライザ）1966
    - パターンマッチだが意外とそれっぽい
  - SHURDLU（シュルドゥル）1972
    - the blue box on the table

自然言語処理の難しさ

- 下線部を漢字に直しなさい
  - このはし渡るべからず

- 以下の文の意味は？
  - 「部長は出かけていなかった」
  部長はいた？いなかった？

自然言語処理の難しさ

- We gave the monkeys the bananas because they were hungry.
- We gave the monkeys the bananas because they were over-ripe.

自然言語処理の難しさ

- We gave the monkeys the bananas because they were hungry.
  → 猿が空腹だったので、バナナをあげた。
- We gave the monkeys the bananas because they were over-ripe.
  →バナナが熟しすぎたので、猿にあげた。
- 最後の単語しか違わないが、theyの差す単語が違う

自然言語処理の難しさ

- 曖昧性
  - 正解が複数考えられる
- 文脈、一般常識
  - 状況や前後の文で意味が変わる
  - 「彼が来るとは思わなかった」
    - 来た？来てない？
  - 「今日暑くない？」
    - （暑い予報だったけど）暑くないよね？
    - （涼しい予報だったけど）暑いよね？
- 創造性
  - 「月が綺麗ですね」（夏目漱石）

自然言語処理の難しさ

- 言語間の多様性
  - 単純に翻訳できない語彙
    - cow/bull/ox/steer/calf/bovine
    - chicken/hen/rooster/chick/broiler
    - 私/わたくし/僕/俺/ワシ/自分、兄弟姉妹
    - 小雨/霧雨/夕立/五月雨
    - 1本/1枚/1匹/1棟/1頭/1等/1投/1答

→そもそも人間にとっても難しい

自然言語処理の歴史

- 自然言語処理の研究開発はコンピュータの開発と同時に開始
  - 1940年代コンピュータの登場とともに、機械翻訳システムの構想が誕生
  - 1950年代情報検索の研究が始まる
  - 1970年代日本でかな漢字変換機能を持つワードプロセッサ（ワープロ）が登場
    - かな漢字変換技術は日本人にとってもっとも身近な自然言語処理の応用技術
    - それまでは辞書を見ながら文字コードで入力したりしていた
  - 1990年代インターネットの普及に伴い、WWW に大量のテキストデータが蓄積される
    - ユーザが欲しい情報を探す情報検索、不要な情報を排除するフィルタリング
    - 膨大なテキストデータの蓄積から「意味のある」または「おもしろい」情報を発掘するテキストマイニング
    - 大量のテキストデータを「コーパス」として蓄積し、コーパスから自然言語処理に用いる知識として規則や確率を獲得できるように

自然言語処理の歴史

- 2000年代　IT 機器の性能向上、大量のデータ蓄積を背景に機械学習手法の適用が始める
  - 入力された言語を別言語へ置き換える機械翻訳が実用的に
  - 人間の発声した言葉を解釈する音声認識が実用的に
  - 第三次AI ブームが開始（ディープラーニング）
    - 画像、音声、自然言語処理分野で従来手法を大きく上まわる性能
- 2010年代　機械学習手法の発展により自然言語処理は飛躍的に進歩
  - RNN やTransformer といったニューラルネットワークモデルが高性能を出す
  - スマートフォンで音声認識による一般的な言葉（文章）での操作が実用的に
- 2020年代　一部の分野では人間と遜色ない（見分けがつかない）レベルへ
  - Transformer を使った大規模言語モデル(LLM : Large Language Model) が注目される
    - GPT (Generative Pre-trained Transformer; OpenAI 社) やGemini (Google DeepMind 社), Claude (Anthropic 社) など
  - 会話用に調整されたAI とチャットができるChatGPT が話題に
    - 自然な質疑応答、テキスト要約、プログラミング言語も扱える、…
    - Microsoft 社やGoogle 社が自然言語処理サービスを拡充

自然言語処理の応用

- 機械翻訳：機械（人間以外）が言語を翻訳
  - Google 翻訳https://translate.google.co.jp/
    - Google 社が提供する翻訳サイト。他にもアプリ版やブラウザ(Google Chrome)統合版なども
  - DeepL https://www.deepl.com/translator
- 文字入力
  - かな漢字変換：入力されたローマ字、あるいはかな文字列を単語に分割して漢字に変換
  - 予測入力：途中まで入力された文字列から、入力したい単語または文字列を予測
  - 入力間違いを補完：多少間違えていても正しく変換
  - なぞり入力/グライド入力

自然言語処理の応用

- 情報検索・検索エンジン
  - Google https://google.com/
  - Yahoo! JAPAN https://www.yahoo.co.jp/
  - Bing https://www.bing.com/
- 対話システム、自然言語インターフェース
  - 人工無脳：特定の単語に反応し、返信するプログラムを揶揄した言い方。知能がないので無脳
    - X(Twitter) で見かけるボット(bot) も同様のものが多い
  - ChatGPT https://chat.openai.com/
    - OpenAI 社の提供するAI チャットサービス
    - API を通じて他のサービスから呼び出すことも可能

- チューリングテスト
  - アラン・チューリングが1950年に提唱したコンピュータの知能テスト
  - 相手がコンピュータであることを隠して、人間とコンピュータを対話させ、人間がコンピュータを人間と誤認したら、コンピュータは知性を持つ、と判定するテスト
  - 当時は文字のチャット、現代ならVRアバターなど
    - ゲームの対戦等でも良い

自然言語処理の応用

- パーソナル・アシスタント：音声ベースでの自然言語によるユーザ補助
  - Apple Siri
  - Google アシスタント
  - Amazon Alexa→Alexa+
- 文章校正、文章作成アシスト
  - DeepL Write https://www.deepl.com/ja/write
  - Grammarly https://www.grammarly.com/
  - Codic https://codic.jp/engine
    - プログラマのためのネーミングサービス
- 文章解析、センチメント分析
  - 迷惑メール（スパム）分類
  - レポートやプログラムコードの盗用検出
  - SNS 投稿の感情（好悪）判定、商品レビューの妥当性判定

自然言語処理の概要

自然言語処理の解析ステップ

- (1) 形態素解析(Morphological Analysis)第3 回
  - 日本語のように、単語間に区切りのない言語では、テキストを単語に分割する
  - 単語が語形変化している場合は原形へ戻す
  - 単語の品詞を決定する
- (2) 構文解析(Syntactic Analysis) 第4,5 回
  - 単語間の構文的関係を決定する
- (3) 意味解析(Semantic Analysis) 第6,7 回
  - 単語、文の意味を決定する
- (4) 文脈解析(Contextual Analysis) 第8 回
  - 複数の文にまたがる処理を行う
  - テキストの構造の解析
  - 照応の解析
  - 省略の補完

自然言語処理を学ぶ意義

- 非常にホットな分野
  - 大規模言語モデル（LLM）の台頭
  - 単に利用するだけではなく、問題を解決し新たな価値を作り出すことが重要
- 分野の理解度が高まる
  - （いわゆるAI でない）従来技術も十分に活用されている
  - 原理や問題点を知ることで、より「上手に」使える
- サービスやAPI を利用すれば良いのでは？
  - 既存サービスを利用するとしても背景知識の有無の違いは大きい
  - 応用・改造には限界がある
- そもそもとして「構文解析ってなんですか？」という人とLLM に関連する仕事をしたいと思うかどうか

まとめ

- 授業全体の進め方を理解する
- 自然言語処理とは、自然言語とは何かを理解する
- 歴史的背景と現代における自然言語処理の応用を理解する