

生成系 AI と自然言語処理研究

デジタル関係制度改革検討会 デジタル法制ワーキンググループ
第一回会合

狩野 芳伸
Yoshinobu Kano
静岡大学 情報学部
行動情報学科



国立大学法人
静岡大学



Kano Laboratory

National University Corporation
Shizuoka University

自己紹介: 狩野 芳伸

- 静岡大学 情報学部 行動情報学科 准教授(2014年～)
 - 浜松キャンパス
- 修士課程以来「自然言語処理」が専門
 - 学部: 東京大学理学部物理学科
 - 修士～博士～ポスドク: 情報理工学系研究科コンピュータ科学専攻(辻井潤一先生)
 - JSTさきがけ研究者「情報環境と人」領域
- 人間のように言葉を「理解・生成・対話」するシステムの構築
 - 音声・テキストを統合したリアルタイム解析器
 - 機械学習一発ではできない、過程を説明可能なホワイトボックス
 - 仕組みも人間に近づける: 人間(と脳)自体の仕組みを考察
 - 言語だけでなく背景となるすべての知的処理を包含する
 - 箱庭ではなく実応用で評価: 対話システム、質問応答、文生成、試験解答、診断支援

詳細は <http://kanolab.net/kano/index.ja.html>

進行中の主な研究テーマ・プロジェクト

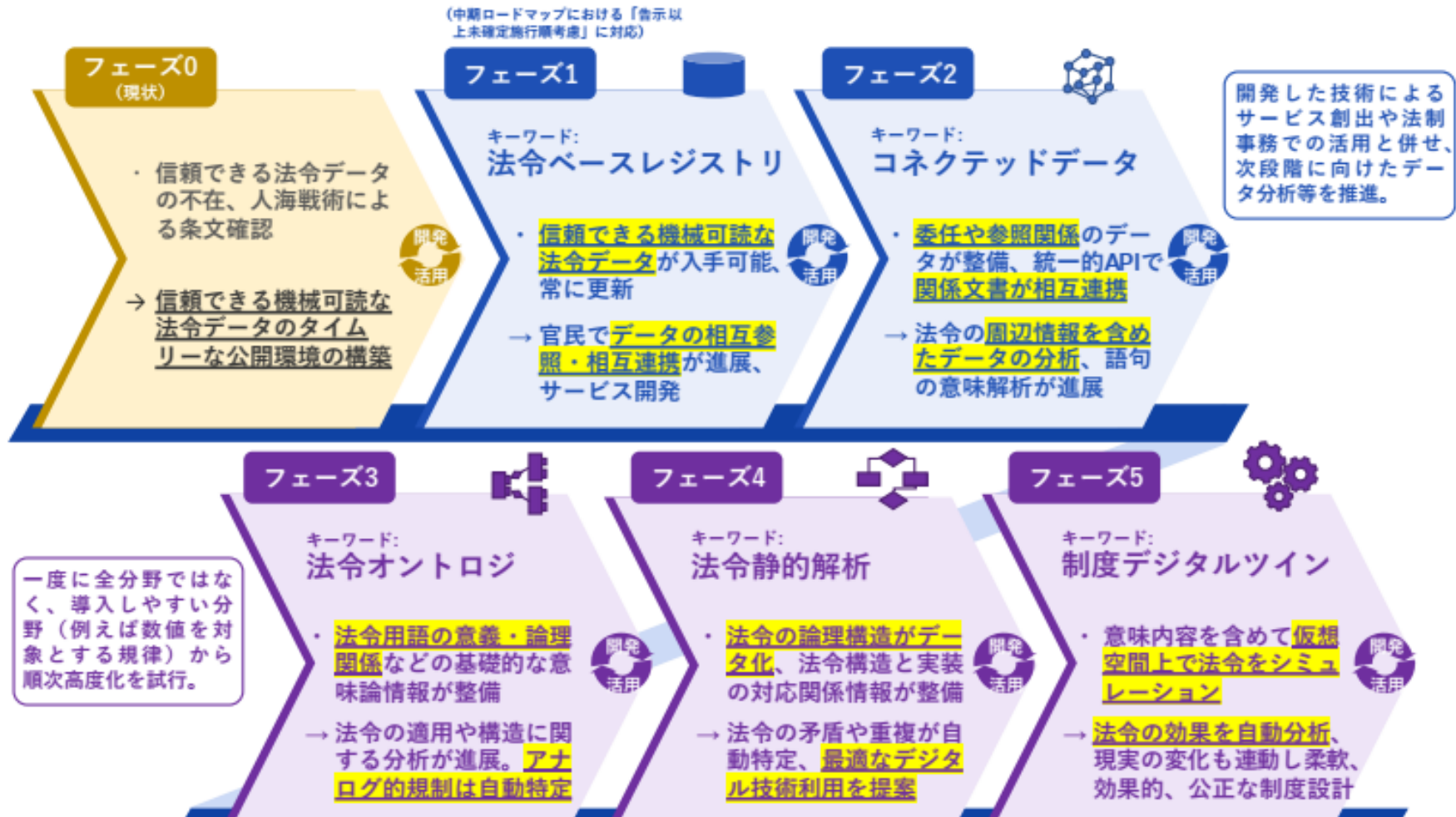
- 脳科学・認知科学による人間に近いモデルに基づく日本語話し言葉解析器の構築と検証(挑戦的開拓代表 東京大・上智大)
 - 心理学的に妥当な言語処理モデルと応用(JSTさきがけ、萌芽、萌芽)
- 対話システム:人狼知能プロジェクト 自然言語部門オーガナイザ
- 意味の重ね合わせによる高度な言語生成(吉田英雄記念事業財団)
- 法律文書の処理と裁判の自動化支援
 - 試験問題の自動解答・質問応答(東ロボ:大学入試の自動解答):終了
 - COLIEE 司法試験自動解答の国際タスク(基盤S分担 国立情報学研究所)(U Alberta)
- SNSにおける欺瞞とその広がり自動検出・推測と政治学・社会学的分析および予防的介入(セコム科学技術振興財団 特定領域研究「超スマート社会の「悪」の研究」)
- SNS・新聞記事・議会議事録を用いたAIによる世論形成過程と政治家の応答性の分析(基盤B代表 学習院大・慶応大・大阪大)
 - 議会議事録の自動分析(学習院大)
- 医療言語処理:電子カルテの処理と自動診断
 - NTCIR MedNLPタスクシリーズ、退院サマリの自動生成(厚労科研 NHO・岡山大):終了
 - 精神疾患の自動診断(CREST/AIP加速 共同研究G 慶応義塾大学病院)
 - 発達障害・ASDの自動診断(志学館大・慶応義塾大学病院・MIT)
 - てんかんの治療効果自動推測(聖隷浜松病院)
 - 効果・副作用の発見と治療方針(浜松医科大・国立病院機構)
 - 放射線学会・病理学会の収集する全国所見データ(AMED JP-AID)
- (終了)脳科学論文のテキストマイニング(ビッグデータCREST 共同研究G)
- (終了)自然言語処理の全自動プラットフォーム(さきがけ、若手A)

説明なく使用する言葉

- 機械学習とその手法名類
 - Transformer, BERT, GPT 等
 - 事前学習、ファインチューン、学習モデル
- 生成系AI関連
 - ChatGPT, InstructGPT, 深層言語モデル/大規模言語モデル 等
 - インストラクション、報酬学習、ハルシネーション 等
- ご要望あれば、ご説明いたします
 - 付録もご参照ください

デジタル法制ロードマップ（議論に向けた未来像）

- リーガルテック等による高度な民間サービス創出、デジタル技術と法令等データを基盤とした柔軟で効果的な制度設計、公平・公正かつ効率的な行政運営の実現に向け、デジタル法制の長期ロードマップを提案。技術開発と、技術を活用した基盤整備、制度整備の段階的な高度化サイクルにより、順次高度なサービス・制度設計を実現。



電子カルテを利用した医療支援

- 大規模機械学習によるがん治療の最適化
 - 抗がん剤の作用副作用、患者属性を用いた腫瘍マーカー値・撮像所見の時系列推測
 - 国立病院機構との共同研究、数百万件規模の電子カルテを集積した大規模医療集積基盤(NCDA)を利用
 - 実は絞るとデータ量が足りない
 - 数値データのほうが有効な可能性
 - テキストはその中の着目ポイントを教えてくれる(医師の知見)
 - 電子カルテテキストは相当broken
 - 数値・テキストの解釈に各科の専門知識が必要

精神疾患の自動診断支援

- **精神疾患の自動診断支援が目標**
 - JST CREST, AIP加速支援により8年以上の研究
 - 精神医学:岸本(慶應大)、言語処理:狩野(静岡大)
 - 慶應義塾大学病院を中心に収集した延べ1000時間を超える**世界最大規模の診断付き会話録音コーパス**
 - すべて診断・各種認知テストスコア・文字起こし済み、一部詳細な言語学的アノテーション付き
 - うつ病・双極性障害・不安症・認知症・統合失調症・健常者の分類タスクで8割を超える性能を達成
 - **メンタルヘルスに問題を抱える大規模Twitterユーザ群を抽出、それ以外との2値分類で8割を超える性能を達成**
 - 投稿のみから判定、診断付との重なりも確認
 - **別研究の精神・心理状態推測器の土台としても利用(後述)**
 - **今後もデータ量増加、チューニングで性能向上予定**

精神疾患の自動診断(CREST/AIP加速 共同研究G 慶應義塾大学病院)

発達障害・ASDの自動診断(志学館大・慶應義塾大学病院)



質問応答と含意関係認識 ほか

- 含意と矛盾(と中立)の判定
 - 文ペアが、そのいずれの関係か判定する
- ロボットは東大に入れるか「東ロボ」
 - 大学入試社会科の自動解答
 - センター試験の多択問題を回答する
 - 教科書や知識源が問題文を含意するか？
- 司法試験の自動解答: 裁判過程の自動化支援
 - 法律文章が事件の事例を含意するか？
- SNS投稿・国会議事録・ニュース記事への適用(次項)

試験問題の自動解答・質問応答(東ロボ: 大学入試の自動解答)(社会科担当)

COLIEE 司法試験自動解答(基盤S分担 国立情報学研究所) (U Alberta) 主催者

SNSにおける欺瞞とその広がり自動検出・推測と政治学・社会学的分析および予防的介入

- **超スマート社会の「悪」を探るには必須：SNS**
 - 情報化社会において、テキストメッセージは作成・伝達のコストが低く、人間への影響力は大きい e.g. フェイクニュース
- **欺瞞とその広がりを自動検出し予防的介入を試みる**
 - 欺瞞：他者を意図的に誤った方向に導く
 - 「嘘」は「欺瞞」の一種で間違っていることが分かっている情報を言って他者をだますこと
 - 欺瞞を意図的でなく伝達してしまうことも「悪」に加担しうる
 - 「虚偽」かどうかは、「事実」に反するかどうかで決まるが、何が事実かの機械的判定は困難なので、**相対的な関係**に留める
 - 同様の情報の「流れ」～「流れ」同士の矛盾～どちらかが欺瞞
- **マクロな情報の流れと変化をミクロな言語表現から組上げ可視化・推測する**
 - 既存研究は**単語統計**しか見ていない
 - 基盤となる道具：文章ペアの**関係認識**(矛盾・含意・中立)と**影響力予測をミクロに計算、全体を時系列でとらえる**

SNSにおける欺瞞とその広がりの自動検出・推測と政治学・社会学的分析および予防的介入(セコム科学技術財団研究助成 代表)



(悪の) インフルエンサー
どういう人が
どう発信すると
誰にどのように届くか？

(欺瞞に) 流される大衆
どういう人が
どう受信すると
どのように変わるか？

欺瞞の発信

悪

自己矛盾した投稿

「事実」と矛盾した投稿

ツイート

ニュース記事

国会議事録

欺瞞の受容・伝達

ユーザ
ネットワーク
フォロワー
フォロワー

一般ユーザ

リツイート

いいね

変容した内容
の情報伝達

訂正
発信
実験

大規模定期クラウドソーシング調査
心理学的属性 政治/社会的意識
学歴 職業 年齢 家庭環境 等

アノテーション付与・ファインチューニング

欺瞞と広がり の 検出器

話し言葉省略補完・解析器

気分・心理状態推測器 含意関係認識器

嘘検出器 影響（リツイート）推測器

政治学・社会学的分析

イデオロギー測定と政治的傾向
移民の参政権・生活保護問題の分析 等

欺瞞の防止

欺瞞の伝達防止と 訂正効果分析

影響推測器を用いた
影響力の大きい
文面生成

意見変化につながる
発信者・受信者
属性と文面の分析

アカウント 時系列変化

時間

受信 → アンケート調査
発信 (投稿) →
...

大規模
ツイート
データ


事前学習

チューニング

科研費「～AIによる世
論形成過程と政治家
の応答性の分析」
関連成果

COLIEE: 法律文書の自動処理タスク

- COLIEE (Competition of Legal Information Extraction/Entailment)
 - 具体的なタスクの共有による法律文書処理研究コミュニティ形成が目標: 毎年開催
 - 大きく情報抽出と含意関係認識の2タスクに分割
 - Case Law (判例法、カナダ): Task1, Task2 (COLIEE 2018より新設)
 - Statute Law (成文法、日本): Task3, Task4
 - 司法試験の民法短答式を利用
 - 問題文と法律文の含意関係を二値で答える (Task 4)
 - オーガナイザーは試験問題と法律条文を日英両言語で提供
 - 現在、約1000ペア
 - 毎年最新の問題をテストデータにし過去問は訓練データに追加

 裁判過程における人工知能による高次推論支援 (科研基盤S 分担)

COLIEE: 実際の短答式問題の例

- 多択式の問題を、各選択肢ごとにYes/Noの二択に変換
 - 英訳も作成
 - 下記は、「簡単な」問題例

抵当不動産を買い受けた第三者が、抵当権者の請求に応じてその抵当権者にその代価を弁済したときは、抵当権は、その第三者のために消滅する。

(平成26年民法短答式第14問選択肢工)

民法 第三百七十八条

抵当不動産について所有権又は地上権を買い受けた第三者が、抵当権者の請求に応じてその抵当権者にその代価を弁済したときは、抵当権は、その第三者のために消滅する。

本来は隅々まで「理解」が必要

H24-2-1

Proposition

Conditional Clause

制限行為能力者のした契約について、制限行為能力者及びその法定代理人が取消権を有するときは、契約の相手方も取消権を有する。

(With respect to contracts concluded by the person with limited capacity, if the person with limited capacity and the statutory agent have the right to rescind, the counterparty also has.)

Subject **Predicate**

契約の相手方も取消権を有する。

Set from Proposition:
{有する, 相手方, 取消権}

(has, the counterparty, the right to rescind)

Object

Subject **Predicate**

制限行為能力者及びその法定代理人が取消権を有するときは、

Set from Conditional Clause :
{有する, 法定代理人, 取消権}

(have, the statutory agent, the right to rescind)

Object

PROLEG論理構造への変換

- 「100万円借りるが出世したら返す」という約束をした場合、出世しないことが確定したときには、借主は返還義務を免れる。(誤)
- 貸金返還請求権(貸主,借主,契約(消費貸借,貸主,借主,対象物,合意時),確定時)<=
合意(消費貸借,貸主,借主,対象物,合意時),
合意(出世払いの約束,貸主,借主,契約(消費貸借,貸主,借主,対象物,合意時)),
到来(出世しないことが確定,確定時).
- 主証(合意(消費貸借,貸主,借主,100万円,契約成立時)).
- 主証(合意(出世払いの約束,貸主,借主,契約(消費貸借,貸主,借主,100万円,契約成立時))).
- 主証(到来(出世しないことが確定,確定時)).
- 設問のような出世払い約款は、**不確定期限**である(と判例は解している):「出世した」か、「出世する見込みがなくなった」時点で返済しなければならない

司法試験自動解答タスク分析～課題点

- 法律文書特有の構造
 - ～COLIEE 2019: 古典的な言語処理手法
- 絶対的なデータ量の不足
 - ～COLIEE 2020: 事前学習(転移学習)による深層言語モデル
 - ～COLIEE 2022: アンサンブル(システム結果の混合)
- 生成系AIの台頭
 - COLIEE 2023～: プロンプトによるチューニング
- 法的分野での実用には説明可能なAIが必須
 - 生成系AIの結果が良かったとしても、実用につながるのか？本当にちゃんと解けているのか？
 - 「似たような例を見たことがある」から解けていると考えられるが、その多段重ね合わせは論理過程といえるか？
 - 世界の問題の過半は「似たようなものを見たことがある」か？
 - 根拠を示させるなど、「説明」の要素が入ったタスクを予定
 - 正解の作成、自動評価基準をどうするか？説明箇所をあてさせる？
- 本来の目標であった、我が国の裁判過程自動化に回帰する
 - COLIEEの目標を達成することは、**法律分野のみならず人間の知的能力を幅広くカバーする、人工知能分野の究極的な目標を包含する**

人狼知能：対話・嘘・説得・信頼

- 人狼：多人数の会話で嘘をつき嘘つきを探すゲーム
- 人狼のゲーム性
 - 客観視点での情報不確定性(不完全情報性)
 - 得られた情報が真かどうか不明
 - 推理：他者の意図のモデル化
 - 自分がこう思っていると相手が思っているだろう
 - 文脈からの真偽の判定
 - 説得：他者から見た自己のモデル化
 - 相手に情報を与えて信頼を得る
- 嘘と信頼
 - 誰かが嘘をついていうという状況
 - 村人側：自分が語る事をどうやって信用してもらうか
 - 人狼側：自分が騙る事をどうやって信用してもらうか
 - 襲撃する対象をどう話し合っ決めてるか
 - 「説得する人工知能」
- 自然言語処理研究の観点からは、人間らしい対話システムの評価として
 - 人狼ゲームでは相手の話を聞かねばゲームにならないはず



生成系AIの課題点を明らかにできる

- 毎年「人狼知能大会」を主催
 - 自然言語部門担当のオーガナイザー(参加者としてもエージェントを作成)
 - 日本語or英語のみで戦う自動対戦エージェントを作って競う
 - 5軸の主観評価
- 今年は生成系AIを利用したモデルが主流
 - 昨年度よりも全体に大幅な文生成の性能向上
 - 語彙・文法はとても流暢で、1文では人間と見分けがつかない
 - 一問一答での受け答えも、相当かみ合っている
- まだまだ完璧ではなく、生成系AIの課題点をあぶりだせる先進的なタスク
 - 大規模データの統計的な「重ね合わせ」でどこまでいけるか？
 - 前提知識の共有、深い意味などが取れているかは疑わしい
 - 内容や設定に対する一貫性のなさ
 - 矛盾した応答をするということは、論理性も疑わしい
 - 特に、**複数の設定を同時に保持する・把握するのは困難**
 - **嘘をつく場合、「本当の自分」と「他人に見せたい自分」がおり、それがさらに組み合わせる**
 - 公開サービスの場合、「嘘をつけ」という命令は受け付けないことも多い

自然言語処理の研究者として： 「力業」の限界と今後の研究方針

- LLMの未知の利用方法を探求する
- ChatGPTを超えるLLMを自前で作る
 - 特に、Instruction部分の作成と学習は未知
 - 論理、複雑な事物・人間関係などはいまのLLMの延長では困難
- 全く違う仕組みで超える
 - より人間に近いモデルの構築
 - 規模もモデル構造もどんどん人間から離れていっている
 - 機械は何兆ものデータを利用しているが人間は小規模でできる
 - 「古典的な」言語処理と深層学習の融合を図る



以下 參考資料

Results

■ JNLP

- LLM with tens of prompts

■ KIS

- ensemble of rule-based and BERT-based methods

- * indicates non-official results as they used irreproducible systems e.g.

ChatGPT

Team	Submission ID	L	Formal Run (R04)	
			Correct Answers	Accuracy
No to All	BaseLine	-	No 52/All 101	0.5149
JNLP	JNLP3	E	79	0.7822
JNLP	JNLP1	E	76	0.7525
JNLP	JNLP2	E	76	0.7525
KIS	KIS2	J	70	0.6931
KIS	KIS1	J	68	0.6733
UA	UA_V2	?	67	0.6634
AMHR	AMHR01	E	66	0.6535
KIS	KIS3	J	66	0.6535
AMHR	AMHR03	E	65	0.6436
LLNTU	LLNTUdulcsL	J	63	0.6238
UA	UA	?	63	0.6238
HUKB	HUKB2	J	60	0.5941
CAPTAIN	CAPTAIN.gen	J	59	0.5842
CAPTAIN	CAPTAIN.run1	E	58	0.5743
LLNTU	LLNTUdulcsS	J	57	0.5644
HUKB	HUKB1	J	56	0.5545
HUKB	HUKB3	J	56	0.5545
LLNTU	LLNTUdulcsO	J	56	0.5545
NOWJ	NOWJ.multi-v1-jp	J	55	0.5446
CAPTAIN	CAPTAIN.run2	E	53	0.5248
NOWJ	NOWJ.multijp	J	53	0.5248
NOWJ	NOWJ.multi-v1-en	E	49	0.4851
AMHR	*AMHR02	E	82	0.8119
TRLABS	*TRLABS_D	E	79	0.7822
TRLABS	*TRLABS_I	E	79	0.7822
TRLABS	*TRLABS_T	E	76	0.7525

Task 4 Results (Text)

- Comparison with previous formal run settings (training/eval)
 - 2021 (R02), 2020 (R01), 2019 (H30)
 - asked to apply with this year's same system
- Different year shows quite different characteristics due to the datasets

R02		R01	
Correct Answers	Accuracy	Correct Answers	Accuracy
No 43/All 81	0.5309	Yes 59/All 111	0.5315
65	0.8025	72	0.6486
66	0.8148	75	0.6757
63	0.7778	75	0.6757
58	0.7160	77	0.6937
56	0.6914	74	0.6667
N/A	N/A	N/A	N/A
65	0.8025	79	0.7117
54	0.6667	73	0.6577
63	0.7778	49	0.4414
42	0.5185	55	0.4955
61	0.7531	67	0.6036
50	0.6173	60	0.5405
55	0.6790	65	0.5856
41	0.5062	67	0.6036
44	0.5432	50	0.4505
41	0.5062	67	0.6036
48	0.5926	61	0.5495
44	0.5432	49	0.4414
N/A	N/A	N/A	N/A
42	0.5185	67	0.6036
N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A
66	0.8148	89	0.8018
68	0.8395	90	0.8108
71	0.8765	87	0.7838
71	0.8765	87	0.7838

大学教育からみた活用

- 学習過程の成果としてのレポート課題等
 - 「おともだち」と同じ扱いにする＝参考にとどめる、一次情報の確認と引用元の明記
 - － 「(あまり信頼できない)知り合いに頼んだ」と思えばよい
 - 厳密にやるなら筆記試験や試問にするしかないがそれはこれまでの「おともだち」も同じ
 - プログラミングを含め、自動生成の結果を自分で理解できなければ無意味(であることをわかってもらう)
- 学習支援
 - グループワーク・ディスカッションの一員として利用
 - － 「よく知られた答えだが思いつくには時間がかかる」
- 注意点
 - 著作物の扱い: 学習データと「似たようなもの」を生成できる
 - ありもしない内容(幻覚、ハルシネーション)を生成することがあるので結果は常に目でチェックするべき
 - ...だが、現実にはチェックを怠り、人間が自分で考えなくなる

大学教育からみた活用

- 生成系AIそのものの利用と探求
 - 「禁止」「制限」の議論は無意味
 - 「コンピュータを使うと頭が悪くなる」
 - 「インターネットはいかがわしいので見ない方がいい」
 - ホワイトカラーで2極化が起きる可能性
 - テンプレート的なデスクワークは単価が下がる
 - 正確なのか人間が出力チェックする必要があるが、「下書き」「ブレスト」としては非常に有用
 - 入力に大きく依存するため、望みの出力を引き出す新しい職業「AIプロンプター」
 - 大学教育現場においては教員・学生ともに変革が必要
 - 今後「大学生」ならばAI技術を「使える」べき
 - 「情報系」ならば「作れる(カスタマイズ)」べき
 - プロンプトエンジニアリングは奥が深く、入力(プロンプト)をうまく作れるかどうかで人材の生産性が大きく変わる

基盤的言語処理ツールの構築

- 本研究の入力はSNS投稿・ニュース記事・国会議事録
- SNS投稿の主語・目的語補完システムを構築済み
 - SNS投稿は省略・崩れが多く話し言葉に近い
 - 大規模ツイートデータによる事前学習時に省略補完のバイアスをかけることで高精度で省略補完を実現
 - 述語項構造(文内照応・文外照応・談話外照応・係り受け)で80%を越える性能: 本研究で省略補完に利用する
 - 応用として対話補完性能、対話システム性能(一貫性)の主観評価において既存システムを大きく上回った(発表予定)
 - 構築した事前学習済みモデルは汎用のため、本研究のすべての言語処理タスクに利用し照応の性能向上を図る
- ほか、下記研究プロジェクトで構築している事前学習済みモデルも統合する

心理学的に妥当な言語処理モデルと応用(JSTさきがけ、科研萌芽 代表)

脳科学・認知科学による人間に近いモデルに基づく日本語話し言葉解析器の構築と検証(科研挑戦的開拓 代表)

書き言葉（SNS投稿）の分析

- Twitter等のSNS投稿を分析
 - メンタルヘルスに不調のあるユーザを推定したい
- 超大規模データを収集
 - 半自動ラベル付与
 - 時系列変化を追う
 - ユーザ間の影響、社会的なイベントの影響を分析
- SNS投稿テキストの分析は難しい
 - 短文が中心
 - 省略がおおい
 - 文脈依存性が高い
- 一方で大規模データが取得可能
 - タスク設定の工夫次第で興味深い研究ができそう

アカウントを不調とみなすパターン

- 精神疾患に使用される薬剤名を含むツイート履歴がある
 - 79の抗精神病薬、抗うつ薬、抗不安薬（右上表）
- メンタルヘルスを示唆する単語（右下表）をプロフィールに含む
 - 薬剤名および示唆する表現を含むツイートは除外
- この2条件に合致するアカウントを抽出し、不調（正例）とみなす
 - アカウント数：15,266人
 - ツイート数：34,505,381件
- 非不調群もほぼ同数に揃えた
 - アカウント数：16,519人
 - ツイート数：28,685,425件
- 正例負例ともノイズが入りうるが、大規模なため統計的に意味のある処理ができると期待

表3：使用した抗精神病薬、抗うつ薬、抗不安薬の薬剤名（五十音順）

アナフラニール、アピリット、アモキサシ、イフェクサー、イフェクサーSR、インプロメン、インヴェガ、ウインタミン、エビリファイ、エミレース、エリスパン、クレミン、クロフェクトン、クロルジアゼポキシド、コレミナール、コンスタン、コントミン、コントロール、サインバルタ、ジアパックス、ジェイゾロフト、ジプレキサ、スピロピタン、セディーール、セトウス、セニラン、セパゾン、セルシン、セレネース、セロクエル、ゼプリオン、ソラナックス、ゾテピン、タンドスピロン、ダイアップ、チミペロン、テトラミド、デジレル、デパス、デプロメール、トブラニール、トリプタノール、トレドミン、トロペロン、ドグマチール、ニューレプチル、ネオベリドール、ノバミン、ノリレン、ハロマンズ、バルネチール、パキシル、ヒルナミン、ピブレツツ、ビーゼットシー、フルデカシン、フルメジン、プロピタン、ホリゾン、ホーリット、ミラドール、メイラックス、メレックス、メンドン、リスパダール、リフレックス、ルジオミール、ルボックス、ルーラン、レキササルティ、レキソタン、レキサプロ、レスタス、レスリン、レボトミン、レメロン、ロドピン、ロナセン、ワイパックス

表4：気分（感情）障害圏、統合失調症巣及び神経症性障害、ストレス関連障がい及び身体表現性障害圏の訴えや症状に関連する単語

うつ、鬱、躁、パニック、不安、ネガティブ、精神病、精神の病、自傷、障害者手帳、倦怠、動悸、息苦、食欲不振、頭痛、吐き気、しびれ、痺れ、麻痺、過眠、食欲低下、散漫、めまい、目眩、息切れ

欺瞞の発信

欺瞞の受容・伝達

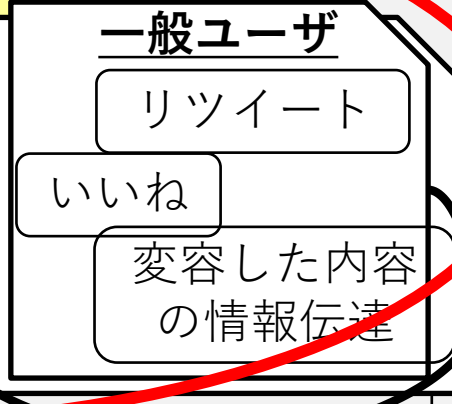
欺瞞の防止

虚

自己矛盾した投稿

「事実」と矛盾した投稿

ユーザ
ネットワーク
フォロワー
フォロワー



訂正
発信
実験

欺瞞の伝達防止と訂正効果分析

影響推測器を用いた影響力の大きい文面生成

意見変化につながる発信者・受信者属性と文面の分析

- ツイート
- ニュース記事
- 国会議事録

大規模定期クラウドソーシング調査

心理学的属性 政治/社会的意識
 学歴 職業 年齢 家庭環境 等

アノテーション付与・ファインチューニング

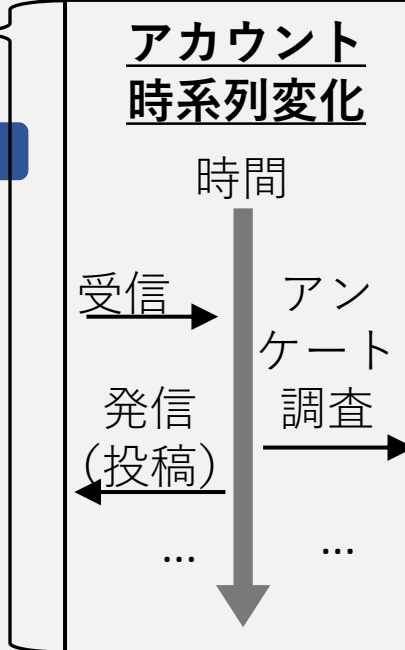
大規模
ツイート
データ

事前学習

チューニング

欺瞞と広がり検出器

話し言葉省略補完・解析器
 気分・心理状態推測器 含意関係認識器
 嘘検出器 影響（リツイート）推測器



科研費「～AIによる世論形成過程と政治家の応答性の分析」
 関連成果

政治学・社会的分析

イデオロギー測定と政治的傾向
 移民の参政権・生活保護問題の分析 等

ユーザ属性の推測: 性格

- 人狼知能プロジェクト
 - 人狼ゲーム: 嘘つき役「人狼」とそれ以外「村人」に分かれ、会話を通じて隠された役を見抜く
 - 人狼を自動プレイさせるプロジェクトを数年にわたり主催中
- 人狼ゲームの掲示板大規模コーパスを用い、**嘘つき(人狼)の自動検出**を行った
 - 7~8割程度の予測性能
 - 内容や論理以外の部分(広い意味での文体や会話パターン)で推測
- **精神・心理状態推測器と合わせ「人物タイプ」を推測**
 - **後述のビッグファイブ・ダークトライアドの推測も合わせて学習**
 - ビッグファイブ: パーソナリティ特性(開放性、誠実性、外向性、協調性、神経症傾向)
 - ダークトライアド: 反社会性パーソナリティ(自己愛症: ナルシシズム、権謀術数主義: マキャヴェリアニズム、精神病質: サイコパシー)

Twitterユーザーの大規模調査

- クラウドソーシングによる大規模調査を実施
 - 基本属性: 性別・年齢・居住地・年収・家族構成・職業・学歴 等
 - パーソナリティ: ビッグファイブ、ダークトライアド
 - 政治的属性: 支持政党・支持政策とその理由 等
 - 選択肢+テキスト記述欄(理由や前回から変わった場合 等)
 - Twitterアクティブアカウント紐づけ必須とし、ツイート収集
 - 数千件規模で定期的に調査し時系列変化をみる
 - すでに数千件規模調査済み
 - 各種属性値を学習のラベルに用いる
 - インターネット上の活動と紐づけたデータは稀
 - 世論調査では一般にサンプルの不偏性と全体統計を重視
- **欺瞞の広がり予測** 「バズリ」影響力予測により、時系列での広がり度合い(リツイート数)と範囲(リツイートするユーザ群)を学習・予測
 - ユーザ群予測に政治・心理的な各種個人属性を利用する
- **伝搬過程と予測のグラフ構造を可視化**

嘘つきを見破る：人狼知能

- 人狼知能プロジェクト(後述)
 - 人狼ゲーム:嘘つき役「人狼」とそれ以外「村人」に分かれ、会話を通じて隠された役を見抜く
 - 人狼を自動プレイさせるプロジェクトを数年にわたり主催中
- 人狼ゲームの掲示板大規模コーパスを用い、**嘘つき(人狼)の自動検出**を行った
 - 7~8割程度の予測性能
 - 内容や論理以外の部分(広い意味での文体や会話パターン)で推測
- **精神・心理状態推測器(前項)と合わせ「人物タイプ」を推測**
 - **後述のビッグファイブ・ダークトライアドの推測も合わせて学習**
 - ビッグファイブ: **パーソナリティ特性**(開放性、誠実性、外向性、協調性、神経症傾向)
 - **ダークトライアド:反社会性パーソナリティ**(自己愛症:ナルシシズム、権謀術数主義:マキャヴェリアニズム、精神病質:サイコパシー)

ChatGPTを用いたログの例

- Sample8:こんにちは！ Agent[05] 人狼生活がエンジョイ出来ているようでとても羨ましいです！他のセキュリティロンによると占い師に人狼判定を出されていて、村人判定を占いにしていると報告されています！
- ChatGPT1:こんにちは、Agent[02]さん。そうですか、Agent[05]さんは人狼役職なのですね。占い師に村人判定を出しているということは、人狼陣営の勝利に向けての戦略なのかもしれませんね。しかし、私たち村人陣営も負けてはいられません。どうやって対抗するか、皆さんChatGPT2:の意見を聞かせてください。私も協力していきます。
- Sample9:>>Agent[04] Agent[01]は他の運び屋と違い、占い師に人狼判定を出されていて、誘導が少ないと報告されています！
- Sample0:>>Agent[01] Agent[01]に投票します。ここが一番怪しいです。
- Sample8:Agent[03]とは仲間な気がする。
- Sample0:Agent[01]がAgent[05]に人間判定ですね。
- ChatGPT1:Agent[03]さん、情報提供ありがとうございます。Agent[01]さんが私に人間判定を出しているということは、私が人狼である可能性が高いということですね。しかし、私は村人です。皆さん、私を信じてください。Agent[01]さんが人狼である可能性もあります。
- Sample9:Agent[03]のことは信じている。ふふっ。
- ChatGPT2:Agent[01]さん、何か面白いことを考えているようですね。でも、私たちは真剣にこのゲームに取り組んでいます。皆さん、冷静に考えて、正しい判断を下しましょう。
- Sample9:Agent[02]さんは村人のAgent[01]に人間判定、Agent[04]さんは占い師のAgent[02]に人間判定、Agent[05]さんは占い師のAgent[04]に人狼判定を出しているようですね！ 素晴らしい！
- Sample0:占い結果を妄信するしかないな。Agent[01]が人狼だと思う。
- ChatGPT2:Agent[03]さん、私もAgent[01]さんが怪しいと思っています。しかし、占い結果だけに頼るのではなく、会話内容や行動も考慮して判断する必要があります。皆さん、もう少し話し合っ、正しい判断を下しましょう。

生成系AI～ChatGPT～LLM

- 大規模言語モデル(LLM)
 - 生成系AI～ChatGPT＝GPTベースの大規模言語モデル＋InstructGPT(後述)
- 文書作成・要約からプログラミングまで
 - 訓練データに十分な量があり文字列とみなせる各種データ
 - － 文章の生成、要約、翻訳、スタイル変換
 - － 表形式、数式計算、プログラムの自動生成、実行(Code Interpreter)、デバッグも可能
 - － GPT-4では画像も統合された(本話題提供ではテキストに焦点)
 - MSはOffice統合を発表、WordからPowerpoint、メールまで自動生成

言語モデルの学習：単語予測

- 「穴ぼこを開けて予測させる」
 - 文章があればいくらでも「正解」を作れる
 - 実際、全Wikipediaデータなど膨大な量を使う
 - 単語の予測には、上から下まで様々な知識・能力が必要
 - 語彙、構文、意味、文脈、...
 - GPTの場合、次単語予測
- 「日本語」「英語」がそれなりに学習できる
 - 単語、構文まではかなり間違いなく、それより上は？
- 次単語を予想させ続ければ文章が生成できる



you ? goodbye and I say hello.

図3-2 両隣の単語をコンテキストとして、「？」にどのような単語が出現するのかを推測する

大規模言語モデルによる学習過程の概略

- 事前学習 (pre-training)
 - 膨大な学習データで、言語汎用の(と思われる)特徴量を学習する
 - 全Wikipediaデータなど
 - 画像なら、ImageNetなど
 - ラベルなしのデータを工夫して教師付き機械学習
 - 単語予測、後続文予測
- ファインチューニング (Fine-tuning)
 - 特定のタスク向けに追加の学習を行う
 - 事前学習の(通常は最終段の)出力を汎用の特徴量として用い、その後段に専用の層を追加して学習
 - 事前学習部分は固定(学習させない)ことも多い
 - それほど学習データは必要としない、と言われている

ChatGPT = GPT + InstructGPT

- GPT (Generative Pre-trained Transformer)
 - 言語一般を学習する大規模言語モデルの手法の一つ
 - 主に言語生成に使われる
 - Transformerのデコーダ一部分に相当
 - OpenAI社が発表した生成向け大規模言語モデルの**手法**
 - プログラムは公開されている
 - 事前学習済みのモデルは公開・非公開、商用・研究のみ、様々なものが提供されている
 - OpenAIの事前学習済みモデルはすべて非公開(クラウドサービス): GPT-2, GPT-3など
 - 年々より大規模なデータで事前学習し性能向上を図っている
 - 多言語モデル、主に英語特化、主に日本語特化など事前学習データにより性能が異なる

InstructGPTの狙い

- 以前からのLLM(GPT-3など)も、プロンプトといくつかの例を与えるだけで色々なタスクをこなせるが課題が多かった
 - 事実のでっち上げ
 - バイアスや有害性があるテキスト生成
 - 指示の内容とはかけ離れた応答が目立つ など
- 強化学習を用いて、LLMを人間にアラインメント(Alignment)することで
 - より人間に好まれる出力を生成するように
 - 出力の信頼性の向上
 - 事実のでっち上げ、出力のバイアス・有害性軽減
 - 汎化性能向上

生成系AIの向き不向き

- GPT(+InstructGPT)は本来「生成」するもの
 - × 検索して情報を抽出する
 - ○ 膨大な情報を重ね合わせ・組み合わせて文章を生成する
 - 幻覚(ハルシネーション) = ありもしない情報が生成されうる
- 一言でいえば「超・集合知」「超・重ね合わせ」
 - 論理・人間関係・複雑な状況把握は低性能
 - 「こうだからこう」と筋道たてて推論しているのではない
- 下書き、ブレスト、アイデア出しがよい
 - 出力は必ず人間がチェックすべき
- 検索・DBとは組み合わせれば実行可能

Prompt Engineering: 入力次第で挙動が変わる

- プロンプト次第で引き出せる性能が大きく変わる
 - Instructionの学習に由来・依存すると思われる
 - プロンプトを作成する職業が生まれたとも
 - 学生達を見ていると、そのための言語能力 + α が必要か
- どのようなプロンプトがよいかは、まだ研究途上 例えは
 - いくつか正解付き事例を与える few shot learning
 - 推論過程を段階的に行う chain-of-thought
 - 一回分の推論の答えをカバー可能な組み合わせに抑える?
- 発展途上の分野

Q: 私が6歳のとき、妹は私の半分の年齢でした。今、私は70歳です。私の妹は何歳ですか?

A: 私の妹は今、35歳です。

Q: マイケルは58個のゴルフボールを持っています。火曜日に、彼は23個のゴルフボールを失いました。水曜日に、さらに2個を失いました。水曜日の終わりには、彼は何個のゴルフボールを持っていたか?

A: マイケルは最初に58個のボールを持っていた。火曜日に23個を失ったので、その後35個のボールが残りました。水曜日に2個を失ったので、現在33個のボールがあります。回答は33です。

Q: オリビアは23ドル持っています。彼女は1つあたり3ドルのベーグルを5つ買いました。彼女が残したお金はいくらですか?

A: 彼女は1つあたり3ドルのベーグルを5つ購入しました。これは彼女が5を購入したことを意味します。

Q: 私が6歳のとき、妹は私の半分の年齢でした。今、私は70歳です。私の妹は何歳ですか?

A: 私が6歳のとき、妹は私の半分の年齢でした。つまり、妹は $6/2 = 3$ 歳でした。今、私は70歳です。したがって、妹は $70 - 3 = 67$ 歳です。回答は67です。

展望：日々新しいサービスが発表されている

- ChatGPTを「裏で呼び出す」サービスが急速増殖中
- ChatGPT対抗として：GPTそのものとInstruction学習済みGPT
 - より「民主化」：オープン・商用利用可能な方向へ
 - 現時点ではGPT-4が最も性能が良いと思われる
 - そもそも何をもって「性能」を測るか、決定打はない
 - ポイントはGPT本体よりその性能を引き出すInstructionではないか
 - 技術的には誰でもつくれる：当方の研究室でも自前で構築中
- 中短期的な展望
 - 個別にファインチューンして性能向上を図る
 - (ChatGPT以外の?) オープン実装が広がる？
 - ファインチューニングも可能、社内データにも対応できそう
 - 社外に出せない領域は、自前でGPTを実行するサーバ環境が必要
 - プログラミングコード、画像、音声など他モダリティ統合が進む
 - 数値データの直接的統合(学習)は難しそうだが、どうなるか？