言語処理学会第24回 年次大会(NLP2018) 報告会

長岡技術科学大学 自然言語処理研究室
小川 耀一朗

面白いと思った発表

- 世界知識の構造に基づいた談話理解モデル
- CE-CLCNN: Character Encoderを用いたCharacter-level Convolutional Neural Networksによるテキスト分類
- 品詞解析の学習者英語への分野適応

世界知識の構造に基づいた談話理解モデル

概要

テーマ：談話解析
著者：　山田隆弘（JAXA）

- 心理言語学の分野において、談話の聴者が有している世界知識が談話の理解において重要な役割を果たすことが知られている
- しかし、どのような知識がどのように使われるかについては明らかにされていない

- 談話の理解において知識がどのように利用されているかを分析し、聴者が知識をどのように利用して談話を理解するかについてのモデルを構築する

世界知識の構造に基づいた談話理解モデル

典型的な行為に限定

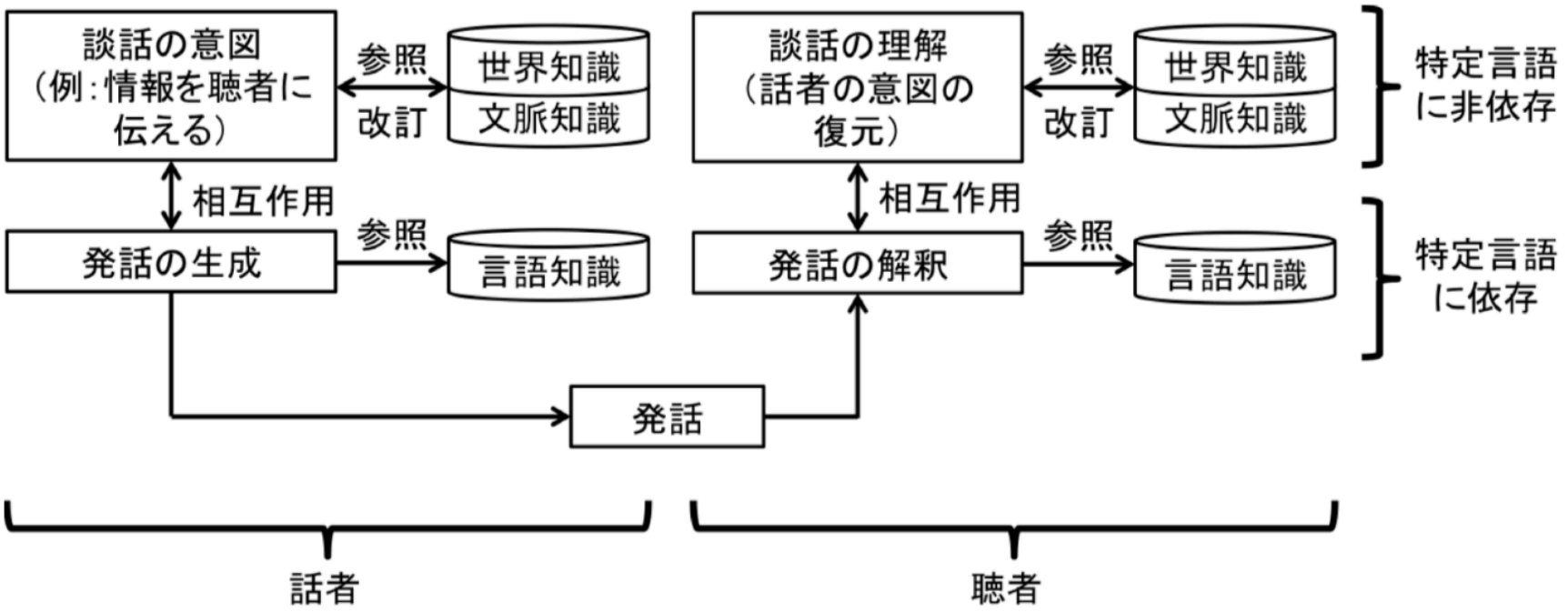
- 最も典型的な言語行為である「聴者の知らない情報を聴者が理解できるように提示すること」を意図した談話の独話を扱うことにする

なぜなら

- “あらゆる言語行為に適用できるモデルを構築することを最終的には目指すべきであるが、初めから完璧なモデルを構築するよりは、まずは典型的な場合から始め、それを徐々に拡張することによって多くの場合に適用できるようにする方が効率的であるからである”

世界知識の構造に基づいた談話理解モデル

知識を用いた談話モデル



- 世界知識：百科事典に書かれているような知識、個人的な情報
- 文脈知識：その談話が行われる状況に依存する知識
- 言語知識：特定の言語の文法と語彙に関する知識

世界知識の構造に基づいた談話理解モデル

所感

- 典型的な行為に限定
  - この考え方は様々なタスクにおいて適用できるのではないか
  - 箱庭言語処理→難易度の低い語彙での行為に限定
  - 誤り訂正→誤りやすい事例(助詞など)に限定

- 知識を用いた談話モデル
  - 談話行為の細分化
  - 3つの知識に対する処理は独立してモデル化できるのはないか

CE-CLCNN: Character Encoderを用いたCharacter-level Convolutional Neural Networksによるテキスト分類

概要

テーマ：文書分類
著者：　北田俊輔, 彌冨仁 (法政大)

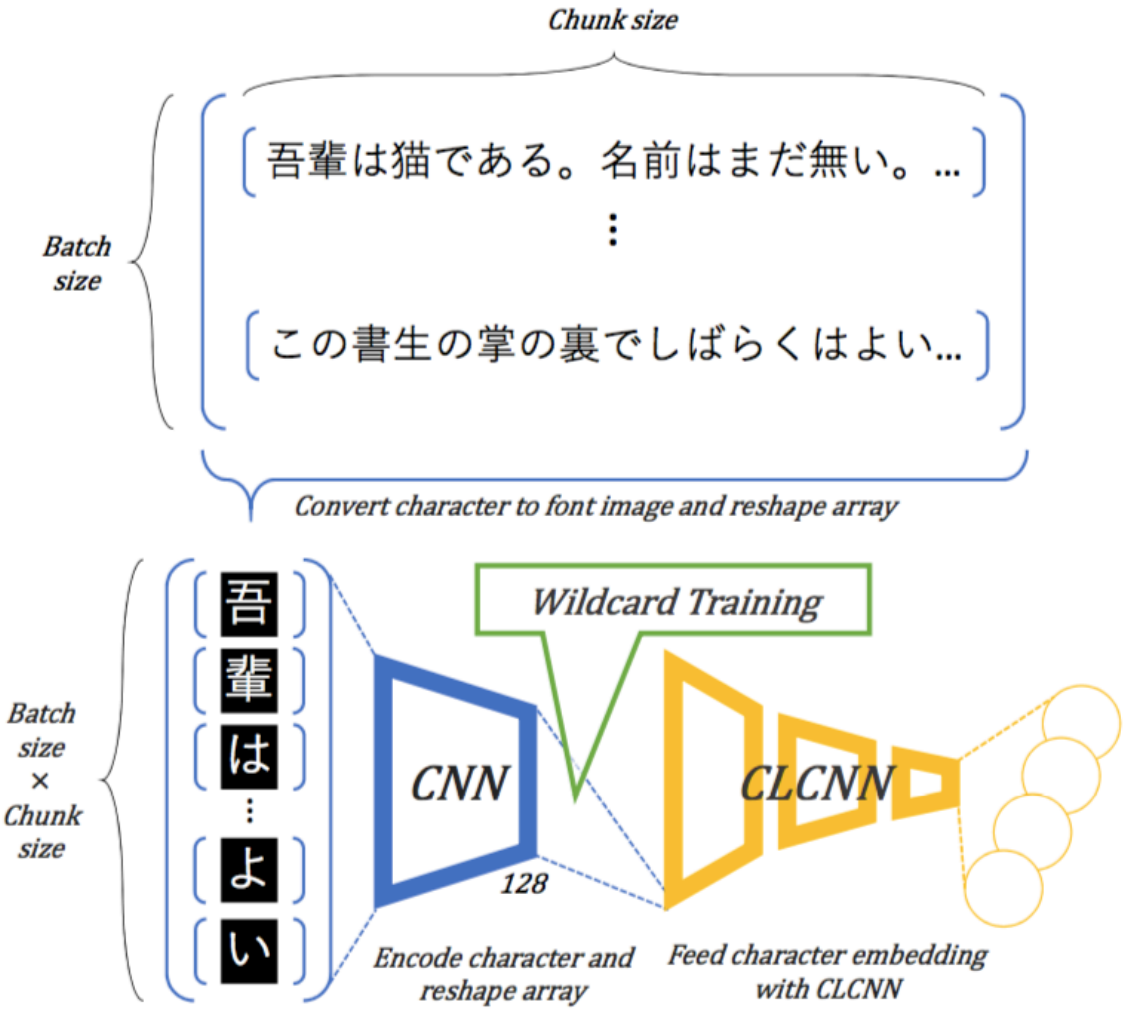
- そもそも日本語の単語分割が難しい
- 文字レベルの特徴を用いた文書分類が提案されているが、文字種の多い言語では過学習を引き起こす

- **入力文の各文字を画像とみなし**、文字埋め込み表現を用いて分類器の学習を行う

- end-to-end て文字表現の獲得から文書分類を行うモデルを提案
- Wikipedia タイトルのカテゴリ推定タスク においてstate-of-the-art の認識精度を実現した

CE-CLCNN: Character Encoderを用いたCharacter-level Convolutional Neural Networksによるテキスト分類

提案手法の全体像とクエリ文字に対する文字表現の近傍上位5文字



クエリ文字	近傍文字	ユークリッド距離
鮫	鰭	370.1
	駮	403.7
	鮪	405.2
	鰐	409.4
	鰤	409.6
痛	癨	317.2
	癪	388.3
	痩	398.3
	痕	398.9
	痴	399.2
披	彼	452.8
	擅	491.5
	擔	520.5
	擒	533.8
	捗	536.8

CE-CLCNN: Character Encoderを用いたCharacter-level Convolutional Neural Networksによるテキスト分類

所感

- 文字の形状に意味がある
  - 表意文字の特徴

- 単語分割が不要
  - 単語分割が難しい誤り文に対して、文字単位で扱うことが効果的か

品詞解析の学習者英語への分野適応

概要

テーマ：言語教育と言語処理の接点
著者：永田亮 (甲南大/さきがけ), 水本智也 (理研AIP), 菊池悠太 (PFN), 川崎義史 (東大), 船越孝太郎 (京大)

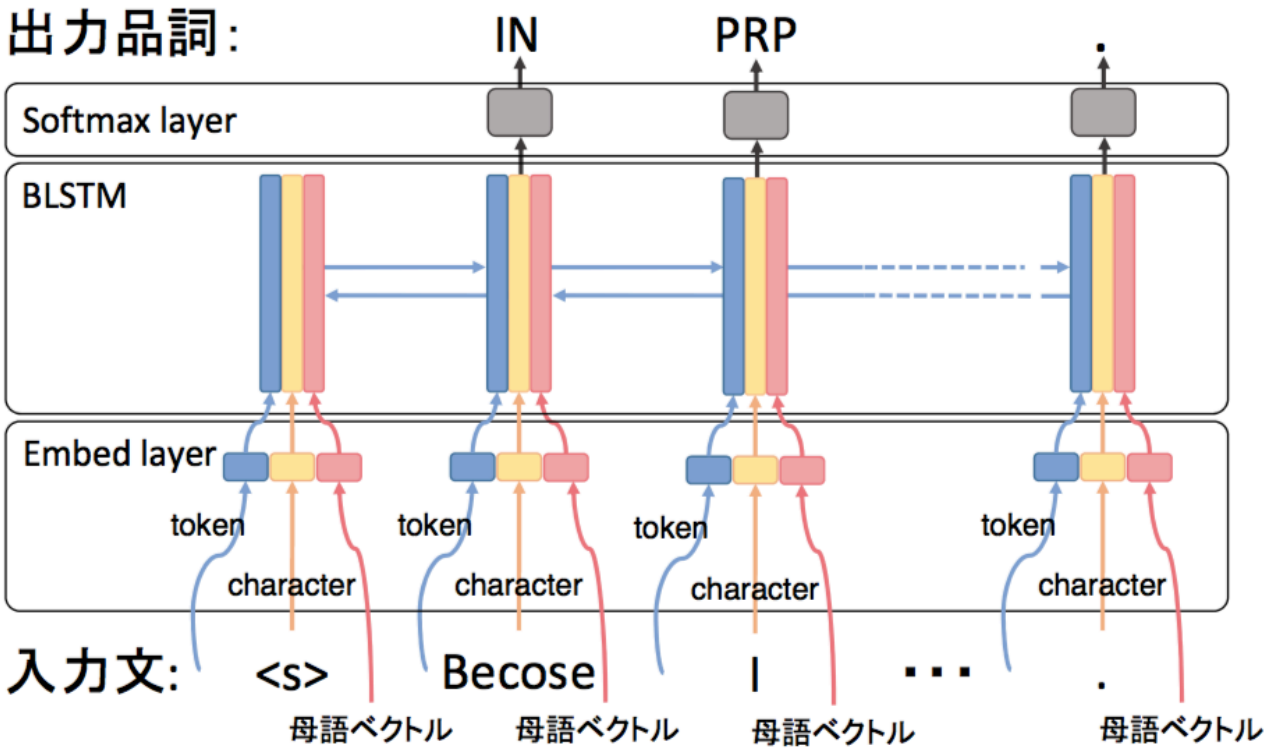
- 学習者英語を扱う研究でも、母語話者用の英文用に開発された品詞解析器を使うことが多い
- しかし、母語話者向けの品詞解析器では学習者の英文を十分な精度て解析てきない可能性かある

*Becose/NNP I/CD like/IN reading/NN ,/,
I/PRP want/VBP many/JJ Books/NNPS ./.

- 品詞解析の学習者英語への分野適合を行なった

品詞解析の学習者英語への分野適応

品詞解析のためのモデル



- **単語自身、単語内の文字列、母語の情報**をベクトルに変換して入力とする
- 単語ベクトルは品詞情報なしの学習者コーパスを母語話者コーパスに加えて事前学習を行うことで、綴り誤りも含めて分散表現の訓練か行われる

品詞解析の学習者英語への分野適応

所感

- 誤りに対してそれっぽい品詞を振る
  - “famouse/JJ”, “thier/PRP$”, “shoud/MD”, “goot/JJ”, “foward/RB”
  - 本来なら”未知語”と判定するべきで違和感があるが、後段の解析にとって必要な考えになるか
- 文字ベクトルや学習者コーパスの単語分散表現を用いる
  - 低頻度であるため分散表現が得られなかった綴り誤りでも正しく解析できる
  - 日本語誤り文に対して文字レベルの情報を扱うことが効果的か
  - （日本語学習者の単語分散表現獲得は、そもそも単語分割がうまくいかないので難しい）

文章校正・誤り訂正に関する発表

- Lang-8を用いた日本語学習者向けの誤用検索システムの構築
  新井美桜, 小平知範, 小町守 (首都大)
- 品詞解析の学習者英語への分野適応
  永田亮 (甲南大/さきがけ), 水本智也 (理研AIP), 菊池悠太 (PFN), 川崎義史 (東大), 船越孝太郎 (京大)
- パイプライン処理によるニューラル英語文法誤り検出と訂正
  金子正弘, 小町守 (首都大)
- 綴り誤りが語彙の豊富さの指標に与える影響の分析
  佐藤太清 (甲南大), 永田亮 (甲南大/理研AIP), 高村大也 (産総研/東工大)
- 誤り文の自動生成による校正エンジンの学習
  中島寛人, 山田剛 (日経イノラボ)
- CBOW言語モデルを用いた契約用語の校正手法
  山腰貴大, 小川泰弘, 中村誠, 外山勝彦 (名大)

面白いと思った公開ツール

- Lang-8を用いた日本語学習者向けの誤用検索システムの構築
  http://cl.sd.tmu.ac.jp/nihongo

- 文献情報の多様な要素を考慮したベクトル表現獲得
  http://tti-coin.jp/demo/bib2vec/

- 論文閲覧を支援する試み ― 文脈検索可能なNLP予稿集コーパス構築
  http://www.mintap.com/nacse/nacse.html