日本語歴史コーパスの語義曖昧性解消の通時適応

東京農工大学
古宮嘉那子

共同研究プロジェクト

- 「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」（代表：浅原正幸）
- 「通時コーパスの構築と日本語史研究の新展開（代表：小木曽智信）
- 科研費 基盤A　日本語歴史コーパスに対する統語・意味情報アノテーション（代表：浅原正幸）

この関係で、古宮研究室の学生さんにやってもらっていた研究です。

（R1年度の田邊絢さんの修論と、）R3年度の多喜凪さんの卒論に相当します。

語義曖昧性解消

- 単語には、意味が２つ以上ある多義語が存在している

手段？

(身体の)
手？

この**手**は有効だ。

- 多義語の語義を判別するタスクを語義曖昧性解消という

翻訳などで重要!!

日本語歴史コーパスの語義曖昧性解消

- 日本語歴史コーパス, CHJは竹取物語、方丈記など日本の古典からなる古文のコーパス

- このCHJに国立国語研究所（国語研）が分類語彙表の語義を付与しつつある（**CHJ-WLSP**）

  分類語彙表：国語研による概念辞書

- 今後古文の新規テキストに語義を自動で付与することを目指し、古文のコーパスの語義曖昧性解消を行いたい

→つまり古文の語義曖昧性解消をしたい

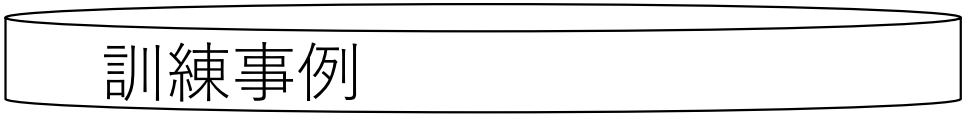
機械学習による語義曖昧性解消

- 文を単語列として考える
- 語義曖昧性解消の対象の単語の前後の文脈から、語義を判定したい
- 「この手を使おう」の「手」は語義1（体の一部）か？それとも語義2（方法）か語義３・・・か？

分類問題の一種

単語ごとに「分類器」を作って判定する（Lexical Sample Task）

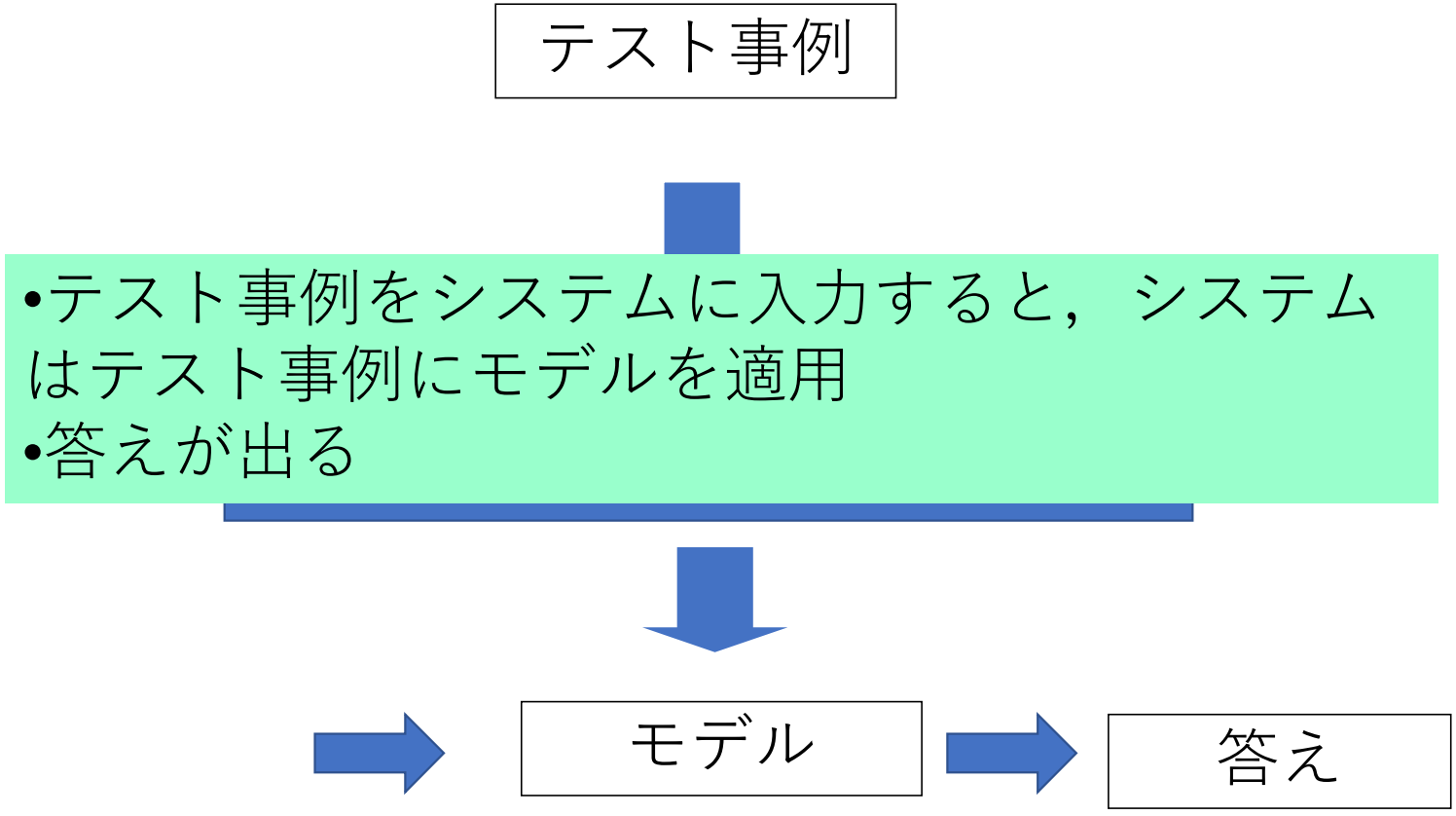
機械学習の枠組み (1) 学習



•訓練事例を入力として，機械学習を行うことで，モデルを作成
•モデルとは，「こういう入力の時にはこうする」というルールの集合
•機械学習では数理的なモデルを**自動的**に作成

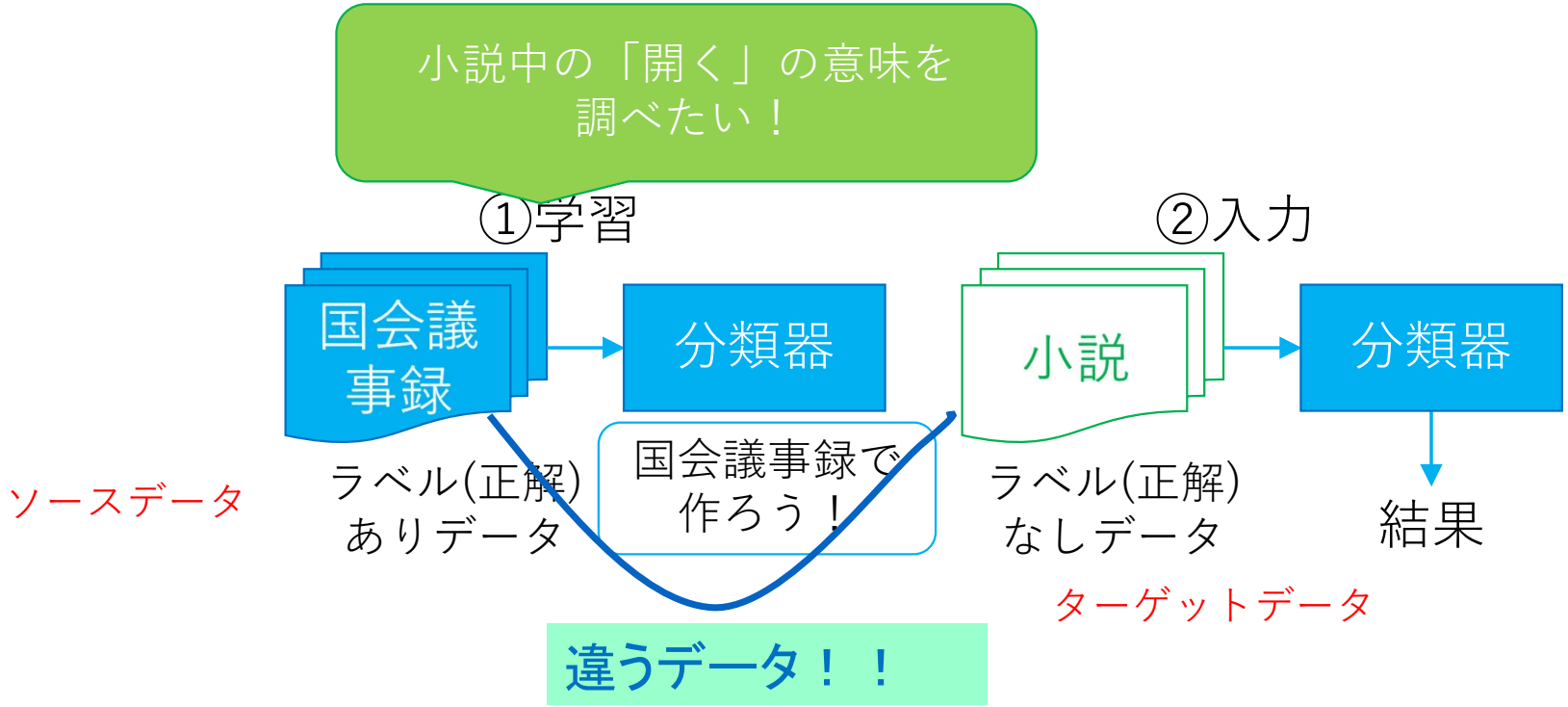
モデル

機械学習の枠組み (2) 推論



ドメイン適応（Domain Adaptation）

- ドメイン適応とは、「学習データとテストデータのドメイン(領域・分野)が異なる際に、学習を適応させる技術のこと
- タグ付けの手間を減らしつつ性能を出したい



古文の語義曖昧性解消のドメイン適応

- 古文の語義タグ付きコーパス(CHJ-WLSP)は量が少ない
- そもそも古文は、タグのないコーパスも、現代文と比べると量が少ない

↓

- 現代文のコーパスならかなりある

↓

ドメイン適応技術を利用する

古文の語義曖昧性解消の通時適応

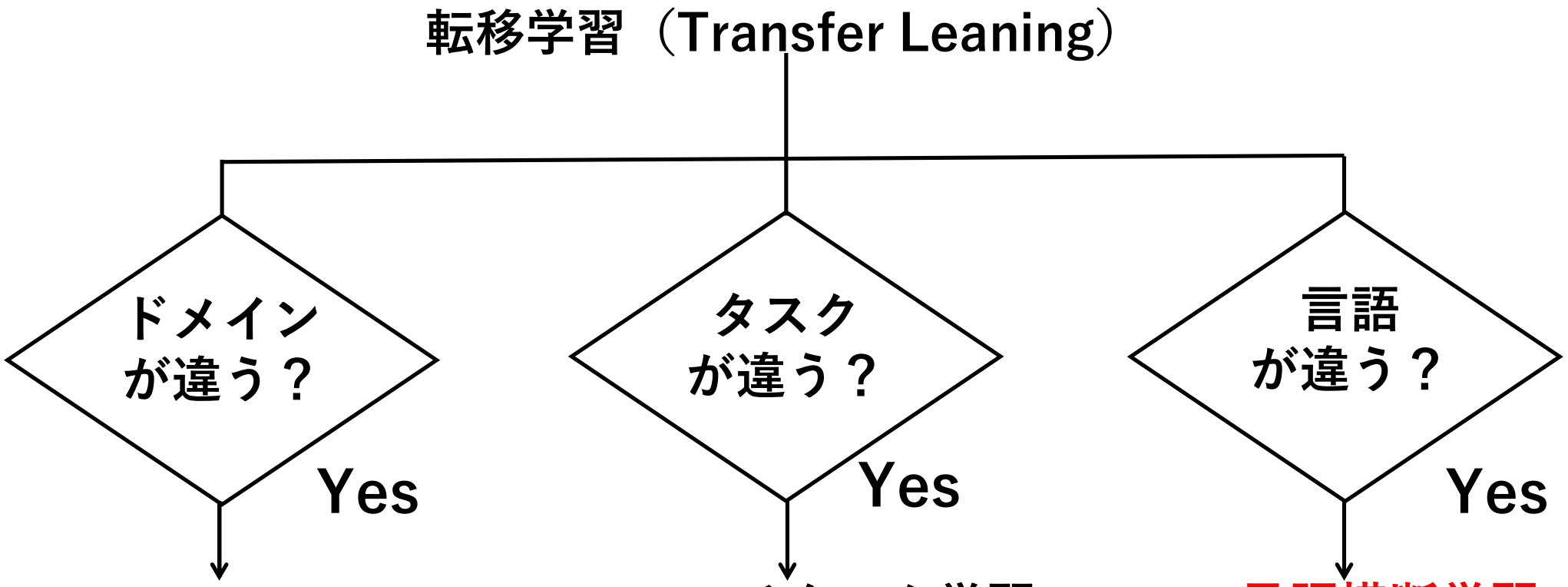
- これまでのドメイン適応→分野Aから分野Bへ

- 本研究のドメイン適応→現代文から古文へ

通時的なドメイン適応→通時適応

(1)分散表現のFine-tuningを利用→R1田邊さんの修論

(2)日本語BERTの Fine-tuningを利用→R3多喜さんの卒論
（今日のTalkのメイン）



通時適応はドメイン適応だが言語横断学習の側面も少しある

- Paul Azunre, Transfer Learning for Natural Language Processing, Manning Publication, (2021)

自然言語処理の
分類器アルゴリズムの変遷

- 自然言語処理の問題の大半が、分類器を作って分類することで解決されている

- 古くは統計的手法（Naïve Bayes）
- 1990年代ごろ？　機械学習
  （長らくSupport Vector Machine（SVM）が主流）
  - 2013年頃～　Deep Learning (Neural Network)
  - 2018年秋に発表　BERT　*Deep Learningの一種
    →事前学習モデルの時代に

自然言語処理の
ドメイン適応技術の変遷

- ドメイン適応技術が脚光を浴びてきたのは2000年代頃
- (DaumeIII, 2007)などが有名

事例ベース：訓練事例を別コーパス（現代文コーパス）から工夫して取ってきて使う

素性ベース：素性（手掛かり、説明変数）を工夫する

・2015年頃～　Deep Learningを利用して分散表現をターゲットデータから作ったり、ソースデータから作った分散表現をターゲットデータを使ってFine-tuningしたりする

・2018年秋に発表　BERT　→事前学習モデルをターゲットデータでFine-tuningする時代に

自然言語処理の
ドメイン適応技術の変遷

- ドメイン適応技術が脚光を浴びてきたのは2000年代頃
- (DaumeIII, 2007)などが有名

事例ベース：訓練事例を別コーパス（現代文コーパス）から工夫して取ってきて使う

素性ベース：素性（手掛かり、説明変数）を工夫する

・2015年頃～　Deep Learningを利用して分散表現を古文データから作ったり、現代文データから作った分散表現を古文データを使ってFine-tuningしたりする

R1　修論

・2018年秋に発表　BERT　→事前学習モデルを古文データでFine-tuningする時代に

R3　卒論

自然言語処理の
ドメイン適応技術の変遷

- ドメイン適応技術が脚光を浴びてきたのは2000年代頃
- (DaumeIII, 2007)などが有名

事例ベース：訓練事例を別コーパス（現代文コーパス）から工夫して取ってきて使う

素性ベース：素性（手掛かり、説明変数）を工夫する

・2015年頃～　Deep Learningを利用して分散表現を古文データから作ったり、現代文データから作った分散表現を古文データを使ってFine-tuningしたりする

R1　修論

・2018年秋に発表　BERT　→事前学習モデルを古文データでFine-tuningする時代に

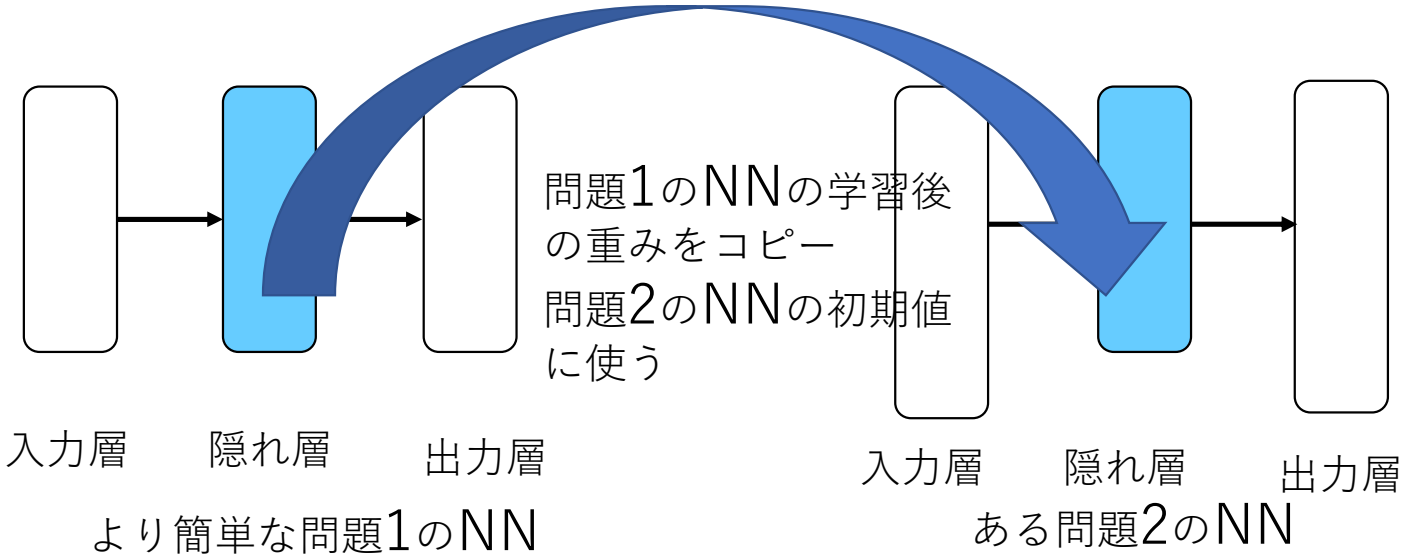
R3　卒論

Fine-tuningとは

- ニューラルネットなどの重み学習の際、その初期値として事前に学習したパラメータを用いること
- 事前の学習に使うデータのタスクと、目的のタスクが似ている場合には、学習される重みはお互いに似ていることが予想できるので、効果が期待できる
- 現代文で学習した値を初期値にして古文に利用すれば役に立つ？

Fine-tuning　（FT）

- ある問題を解くときに、それとよく似た（もっと訓練事例がたくさん手に入る）問題を解いた際の重みを初期値として使う手法



分散表現はNNで作るのでこの手法で通時適応できる！
現代文のデータを利用すればいい

R1時点でのCHJ-WLSP

- 古文のコーパス：日本語歴史コーパス, CHJ

作品名	単語数	時代	スタイル
方丈記	5,402	1212年	随筆
竹取物語	12,757	900年頃	物語
虎明本狂言	5,448	1642年	狂言
土佐日記	8,208	934年	日記
徒然草	40,834	1332年	随筆

- 出現回数が古文でも現代文でも50以上の単語だけを対象にした Lexical Sample Task

語義曖昧性解消の対象語 (58語)

為る	居る	一	言う	有る	来る
成る	様	人	年	日	何
物	時	見る	行く	中	後
月	此れ	思う	良い	所	持つ
方	万	又	今	家	入る
内	上	間	聞く	取る	男
彼	知る	作る	国	置く	付ける
見える	他	女	書く	心	下
共	皆	唯	道	立つ	
身	或る	読む	返る	事	

R1修論　結果

シナリオ	両方		古文のみ		現代文のみ	
素性	マイクロ	マクロ	マイクロ	マクロ	マイクロ	マクロ
古文300素性	63.12	69.56	70.57	74.28	48.20	59.24
古文200素性	63.10	69.41	70.41	73.94	48.51	59.05
現代文素性	66.69	71.27	69.79	73.41	51.21	61.05
古文素性＋現代文でFT	66.27	72.03	**70.80**	**74.83**	50.64	61.75
現代文素性＋古文でFT	66.01	70.45	68.33	72.42	47.08	57.58
現代文素性＋古文で時代順FT	66.27	70.76	69.35	73.39	47.87	56.35

素性作成手法(fine-tuning)

- **古文の素性**：古文だけでWord2Vecを学習した素性
  古文300素性：300次元のもの
  古文200素性：200次元のもの
- **現代文素性**：nwjc2vecを利用
- **古文素性＋現代文でFT**：古文によって学習された素性（300次元）をBCCWJでfine-tuningする
- **現代文素性＋古文でFT**：nwjc2vecを古文でfine-tuningする
- **現代文素性＋古文で時代順FT**：nwjc2vecを古文でfine-tuningする。この際に時代順のサブコーパスによって段々と時代を戻してfine-tuningする

R1修論　一番良かった時の実験設定

- 分類器はSVM
- 素性とする分散表現はword2vecを利用

- 利用した語義タグ付きデータ：CHJ
- 作成済み現代文のword2vec：nwjc2vec（10億文から学習）
- 古文のテキストコーパス：小学館コーパス

タグなしデータ1　分散表現作成用

- 古文のコーパス：小学館コーパス
- 古代から近世までの文学・歌集作品を収録
- 日本霊異記, 古今和歌集, 竹取物語, 伊勢物語, 大和物語, 平中物語,土佐日記, 蜻蛉日記, 落窪物語, 堤中納言物語, 枕草子, 源氏物語,和泉式部日記, 紫式部日記, 更級日記, 讃岐典侍日記, 大鏡, 今昔物語集, 将門記, 陸奥話記, 保元物語, 平治物語, 方丈記, 徒然草, 正法眼蔵随聞記, 歎異抄, 平家物語, 宇治拾遺物語, 十訓抄, 沙石集, 曾我物語, 近松門左衛門集, 洒落本, 滑稽本, 人情本, 俊頼髄脳, 古来風躰抄, 近代秀歌, 詠歌大概, 毎月抄, 国歌八論, 歌意考, 新学異見　義経記　室町物語草子集　謡曲集　一休ばなしなど
- 49作品
- 合計3,468,009単語

タグなしデータ２　分散表現作成用

- 現代文のコーパス：現代日本語書き言葉均衡コーパス(BCCWJ)

ジャンル名	単語数	時代	スタイル
PB	111,983	現代	書籍
PN	117,543	現代	新聞
PM	117,568	現代	雑誌

- 合計121,332,694単語

素性

- 語義曖昧性解消の対象単語の左右2つの単語の分散表現

- 例：

この / 手 / を / 使お / う / 。

対象単語が「手」の場合

赤字の単語の分散表現を連結したものが素性ベクトル

※「この」の前に単語はないのでゼロベクトルを連結する

R3　BERTを利用した古文の語義曖昧性解消

- BERT：東北大のbert-base-japanese-whole-word-masking
  日本語のWikipediaから学習したモデル
  ほぼ現代語から学習していると言えるので、
  通時適応の一形態とみなせる

入力は文ベース
文には語義曖昧性解消の対象語が含まれる
この対象語のBERTの出力ベクトルが最終層の入力→Fine-tuning

R3卒論　結果

Model	Micro Avg.	Macro Avg.
R1修論のMFS	70.00%	75.54%
R1修論の提案手法	70.80%	74.83%
R3卒論のMFS	69.44%	75.70%
BERT利用（提案手法）	73.15%	79.68%

- R1修論では五分割交差検定を行っているが今回は交差検定していない
  →MFS（最頻出語義）が異なっている
- R3卒論は3回ずつ実験した平均値を利用
- 有意に正解率が上昇

→現代文のBERTにより古文の語義曖昧性解消は性能が上がる
（通時適応可能）

R3時点でのCHJ-WLSP

- 追加分（ぐっと量が増えた）

作品名	単語数	時代	スタイル
今昔物語集	175,601	1120年頃	説話集
十訓抄	90,177	1252年？	説話集
宇治拾遺物語	120,705	1221年	説話物語集

- 出現回数が古文のみで500以上の単語だけを対象にしたLexical Sample Task

語義曖昧性解消の対象語 (38語)

為る	知る	行く	取る	下	申す
成る	人	国	心	無い	程
物	見る	返る	立つ	是	参る
居る	思う	有る	事	其	
様	間	日	来る	出でる	
時	女	所	後	然る	
此れ	言う	家	持つ	者	

重み共有の実験

- シンプルなBERTモデルに加えて重み共有も試してみた

(1)語義曖昧性解消の対象単語のFine-tuning
　ひとつめの単語のモデルをふたつめの単語の初期値に
　これを全単語繰り返す
(2)語義曖昧性解消と、その出典のマルチタスク学習

→両方効かなかった

R3卒論　1)語義曖昧性解消の対象単語のFine-tuning

- R1までのデータ

Model	Micro Avg.	Macro Avg.
MFS	71.19%	75.62%
The simple BERT model	**74.34%**	**79.42%**
FT among target words	73.47%	79.17%

- BERTを使うとMFS（最頻出語義）より有意に正解率が上昇
- 重み共有では、有意差はないが少し落ちる

R3卒論　1)語義曖昧性解消の対象単語のFine-tuning

- R3までのデータ

Model	Micro Avg.	Macro Avg.
MFS	62.68%	58.30%
The simple BERT model	**83.30%**	**85.96%**
FT among target words	82.37%	85.49%

- BERTを使うとMFS（最頻出語義）より有意に正解率が上昇
- 重み共有では、有意差はないが少し落ちる

R3卒論
2)語義曖昧性解消と、その出典のマルチタスク学習

- R1までのデータ

Model	Micro Avg.	Macro Avg.
MFS	69.81%	73.79%
The simple BERT model	**72.82%**	**77.50%**
FT among target words	72.55%	77.24%

- BERTを使うとMFS（最頻出語義）より有意に正解率が上昇
- 重み共有では、有意差はないが少し落ちる

R3卒論　R1のデータ　出典

作品	テストデータ数
竹取物語	350
土佐日記	269
方丈記	145
徒然草	1,103
虎明本　狂言	106
合計	1,973

- 徒然草が半数以上を占める

R3卒論
2)語義曖昧性解消と、その出典のマルチタスク学習

- R3までのデータ

Model	Micro Avg.	Macro Avg.
MFS	62.27%	58.35%
The simple BERT model	**83.02%**	**85.75%**
FT among target words	81.82%	85.32%

- BERTを使うとMFS（最頻出語義）より有意に正解率が上昇
- 重み共有では、有意差はないが少し落ちる

R3卒論　R3のデータ　出典

Literature	Number of Test Data
竹取物語	347
土佐日記	208
方丈記	121
徒然草	992
虎明本　狂言	89
今昔物語集	5,091
十訓抄	1,969
宇治拾遺物語	3,280
合計	12,097

- 今昔物語集と宇治拾遺物語が多い

R3卒論
R1のデータとR3のデータの比較

Data	Model	Micro Avg.	Macro Avg.
R1	MFS of test data	69.44%	75.70%
R1	The simple BERT model	73.15%	79.68%
R3	MFS of test data	62.38%	58.24%
R3	The simple BERT model	81.75%	84.93%

- BERTを使うとMFS（最頻出語義）より有意に正解率が上昇
- R3のデータではMFSは下がったが、BERTを使ったモデルの正解率はかなり上がって8割以上に

考察

- R1修論の際はnwjc2vec（10億文コーパスから作成）を利用しているがBERTは1700万文のWikipedia

→事前学習のデータ量が理由で語義曖昧性解消の精度が上がったのではない

- 現代文BERTにおける古文の未知語は

　R1データで18.32%、　R3データで17.06%であったが語義曖昧性解消の正解率に悪影響は（少なくとも恐れていたほどは）なかった

- ドメイン内の訓練事例（Fine-tuning用の訓練事例）の数は

　R1　170.1用例/一単語

　R3　1519.7用例/一単語　　8.93倍

まとめ

- 古文の語義曖昧性解消を現代文のコーパスを使って通時適応した
- BERTをFine-tuningする方法は、既存手法（SVM使用、分散表現のFine-tuning）を優位に上回った→通時適応できている
- 二種類の重み共有は有効ではなかった
  - 語義曖昧性解消の対象単語のFine-tuning
  - 語義曖昧性解消と対象単語の出典の文書分類のマルチタスク学習
- Fine-tuning用の訓練事例が増えることによりかなり性能向上