語義曖昧性解消
コーパスへの意味タグの付与システム

東京農工大学
古宮嘉那子

第 64回 語彙・辞書研究会
2023 年 11 月 18 日

語義曖昧性解消

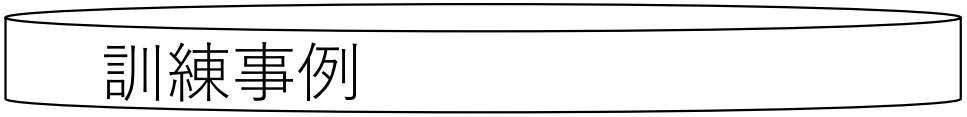
- 単語にはさまざまな意味（＝語義）があり、文脈によって使い分けられている
- この語義を一意に決定するタスクを語義曖昧性解消という

例：「この手を使おう」の「手」

→×体の一部　○方法

- 語義曖昧性解消において、辞書はあらかじめ与えられる

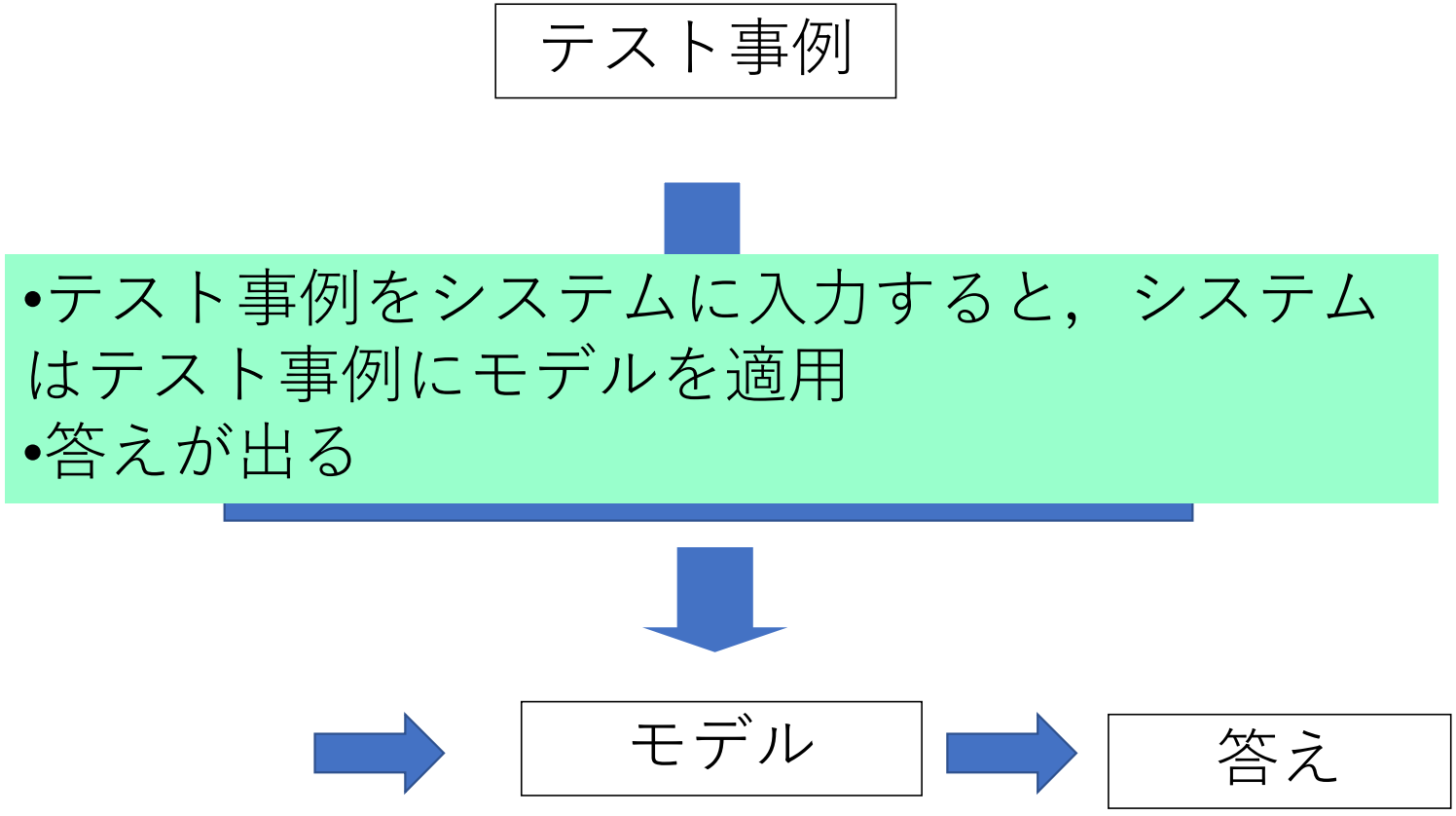
機械学習の枠組み (1) 学習



•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



性能と汎用性

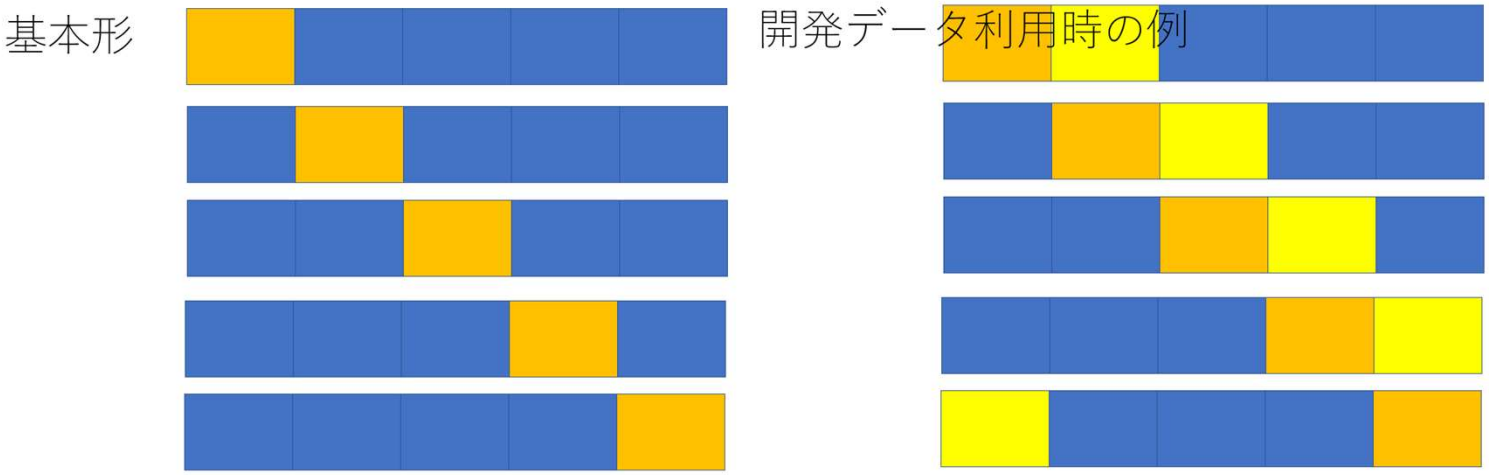
「性能」：正解率など。
「汎用性」：いろいろな問題が解けること。

機械学習の目的は性能を上げることだけではなく、
汎用性をもたせつつ性能を上げること

手持ちの問題集だけ答えられても意味がない
過学習と未学習を防いでちょうどいい学習を
→ハイパーパラメータの選択

自然言語処理の研究イメージ

- データの一部を機械学習の訓練用に（一部、パラメータ決定用）
- 残りを性能のテスト用に利用する



よりよい技術を開発すること自体が研究の目的

タグ付きデータの一部を答え合わせ（テスト）に使う

語義曖昧性解消の種類

1)コーパス中の頻出語のみを対象とする（**Lexical Sample Task**）
　単語ごとに「分類器」を作って判定する

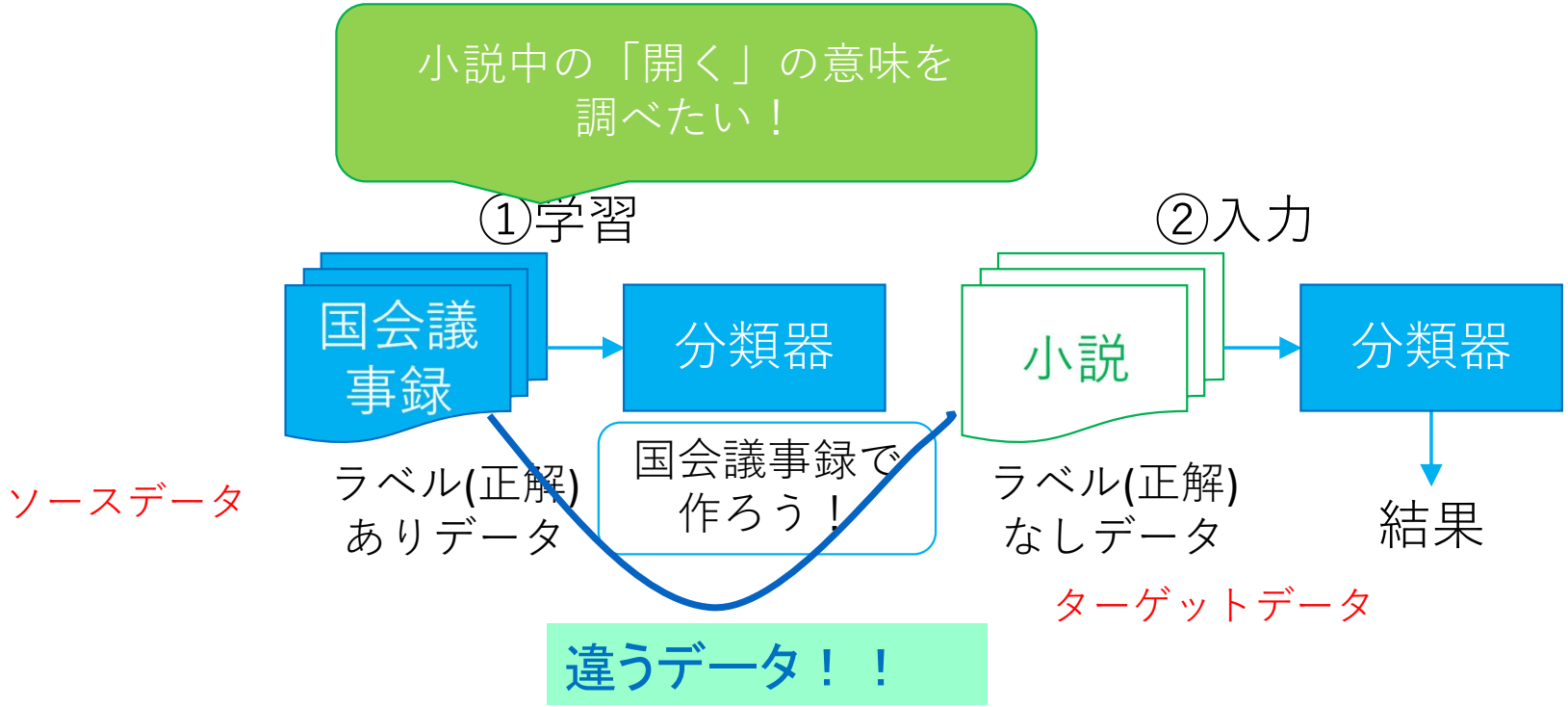
2)コーパス中の全単語を対象とする（**all-words WSD**）
　系列ラベリング（という手法）で解く

用例が少ない単語・用例が少ない語義は難しい
技術の進歩によりより実用的な　2) all-words WSDも研究対象に

ただし、タグ付きコーパスが必要（今のところ内容語のみ）

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



One-hot vector

- テキスト中に出てくる単語タイプ数（語彙数）分のベクトルを作っていた
- ひとつの単語を表すときは、そのうちのひとつだけ1になる

語彙数が７のとき

インデックス番号2の単語を表すone-hot vectorは

(0, 1, 0, 0, 0, 0, 0)

インデックス番号5の単語を表すone-hot vectorは

(0, 0, 0, 0, 1, 0, 0)

しかしこれだと

- 単語を離散的に扱っているため、

「猫」：　　0000001
「犬」：　　0000010
「ディープラーニング」：0000100

「猫」と「犬」、「猫」と「ディープラーニング」の違いは同程度とされてしまう

分布意味論の分布仮説

- 言語学における分布意味論の分布仮説
「文脈の似た単語・句・文などは意味が似ている」

文脈が意味を表す

意味を表したい単語の周辺の単語の頻度

これをベクトルとして表す

Bag of Words (BOW)

- テキスト中に出てくる単語タイプ数（語彙数）分のベクトルを作っていた
- ひとつの単語を表すときは、その周りに出てきた単語の値が1になる

語彙数が７のとき

インデックス番号2,5の単語がある単語の周りにあれば

(0, 1, 0, 0, 1, 0, 0)

インデックス番号7の単語がある単語の周りにあれば

(0, 0, 0, 0, 0, 0, 1)

しかしこれだと（位置を考慮すれば）

- 語彙数×文脈を表すウィンドウサイズ分のベクトル

「猫」：0 0 0 0 0 0 1　0 0 0 0 0 1 0　0 0 0 0 1 0 0…

「犬」：0 0 0 0 0 1 0　0 0 0 0 1 0 0　0 0 0 0 1 0 0…

「ディープラーニング」：0 0 0 0 1 0 0 0 0 0 0 1 0 0…

周辺に出てくる単語しか1以上にならない

→ほぼ0のベクトルになってしまう

データスパースネスの問題（データが疎）

→機械学習（統計処理）にはよろしくない

分散表現

語彙数よりも低次元のベクトルとし、実数値の値をもつ密なベクトル

たとえば、さっきの例の単語は、

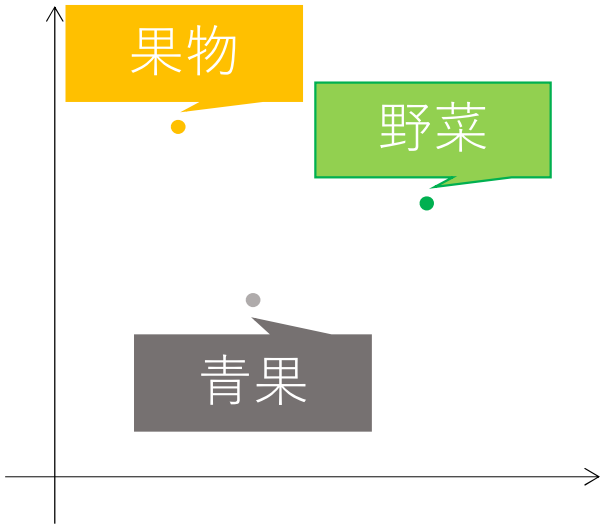
(0.2, 1.3, 0.4, -0.05)　などになる

（四次元とした場合。この次元数は自分で決める。普通は200とか300.）

深層学習による単語の元祖分散表現がWord2Vec（Mikolov 2013a, 2013b）

単語の分散表現

- Word2Vecの出現により、単語の意味をベクトルとして計算できるようになった



意味の似ている単語は、
分散表現も似ている！

そうなるように深層
学習で学習してある

単語の分散表現

- 特に、Word2Vecは構成性（compositionality）をもつ

king-man+woman=queen

London-UK+France=Paris

などの意味の計算ができる

- このように，ある単語の意味が二つ以上の単語の意味から構成できることを構成性をもつという
- Word2Vecの学習アルゴリズムにはCBOWとSkip-gramがある

この時代は「語彙の」分散表現

- その他にも「単語の分散表現」には色々な種類があります

(Mikolov+ 2013a, 2013b)のWord2Vecだけでなく

(Bojanowski+ 2016), (Joulin+ 2016) のfastText

(Pennington+ 2014)のGloVeなどが有名

これらはどれもこれも，「語彙」につきひとつのベクトルができる

語義/句/文の分散表現

- 単語列の代わりに意味でしるしをつけた語義列を使えば，語義の分散表現もできる

- 単語の分散表現の足し算または最大値をとるなどして，句や文の分散表現もある

(句や文の分散表現はそれほど性能がよくないのが現状．多様性に対してデータが足りない)

語彙から出現へ

- (Peters+ 2018)によるELMo　は，LSTMの中間表現そのものを分散表現とした。

  語彙ごとではなく出現ごとに分散表現を作成

→文脈ごとに違う意味ベクトルができる

語彙と出現：

「私は私」という　文で，語彙で数えれば単語は2つあるが，出現で数えると単語は3つである.

この例で見ると，ひとつめの「私」とふたつめの「私」では違うベクトルを作るようになったのがELMo

「意味」は文脈の中にある

ElMo
(Embeddings from Language Models)

- (Peters+ 2018)による出現レベルの分散表現の提案
- LSTMで言語モデル(文脈を入力として次の単語を予測するモデル)を作り、そのネットワークの重みを分散表現とする
- 双方向 LSTMを複数層（論文中では2層）使う
- この隠れ層の重みの重み付き線形和をとる

  →これを分散表現として使う

  下層の隠れ層は文法的、上層の隠れ層はより意味的

BERT　(Bidirectional Encoder Representation from Transformer)

- (Devlin+ 2018)による，ELMoより性能のよい出現レベルの分散表現
- 基本的には12段のTransformer
- 以下のふたつのタスクをとくことで、事前学習し、性能の良い分散表現を得る（ラベルなしデータを利用。）

Masked Language Model

Next Sentence Prediction

- 事前学習の重みを使って、fine-tuning(転移学習の一種)により個別のタスクをラベルありデータで解く

Open AI　GPT (2018)
(Generative Pre-trained Transformer)

- BERTの少し前に提唱された大規模言語モデル
- Improving Language Understanding by Generative Pre-Training
- 次の単語を予測するcausal modeling→言語モデル

大量の生テキストでニューラルネットワークを学習して、そのあと、Fine-tuningをして各タスクを解くことを提唱

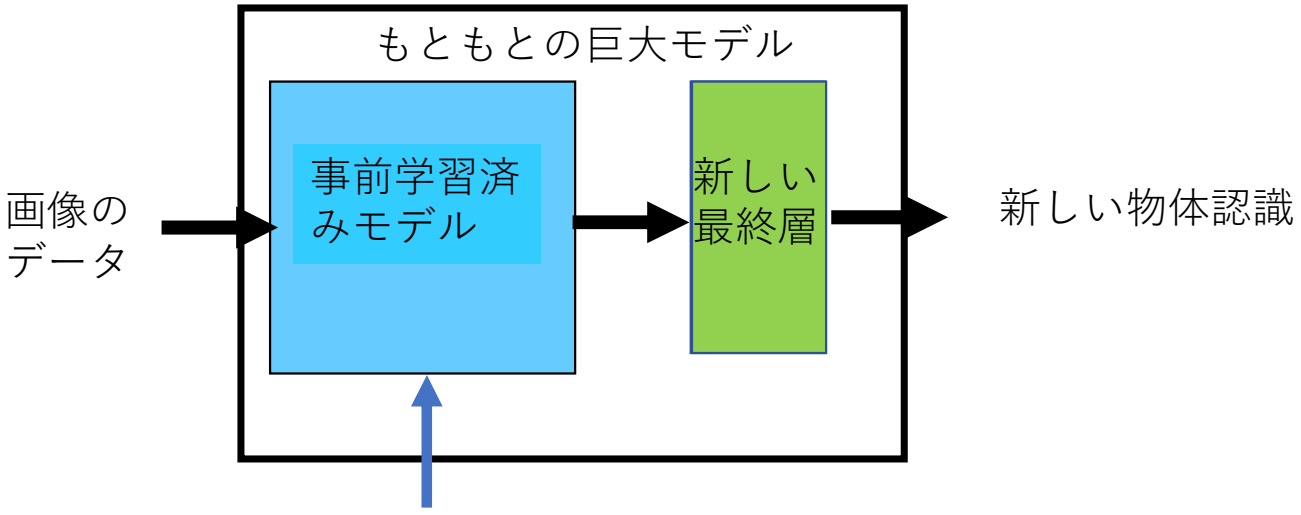
各タスク別にネットワークを作るのではなくて、一般的な事前学習モデルをタスクごとに微調整

→　BERTはこの流れを汲んでいる

大規模言語モデル

- 自然言語処理の事前学習済みモデル

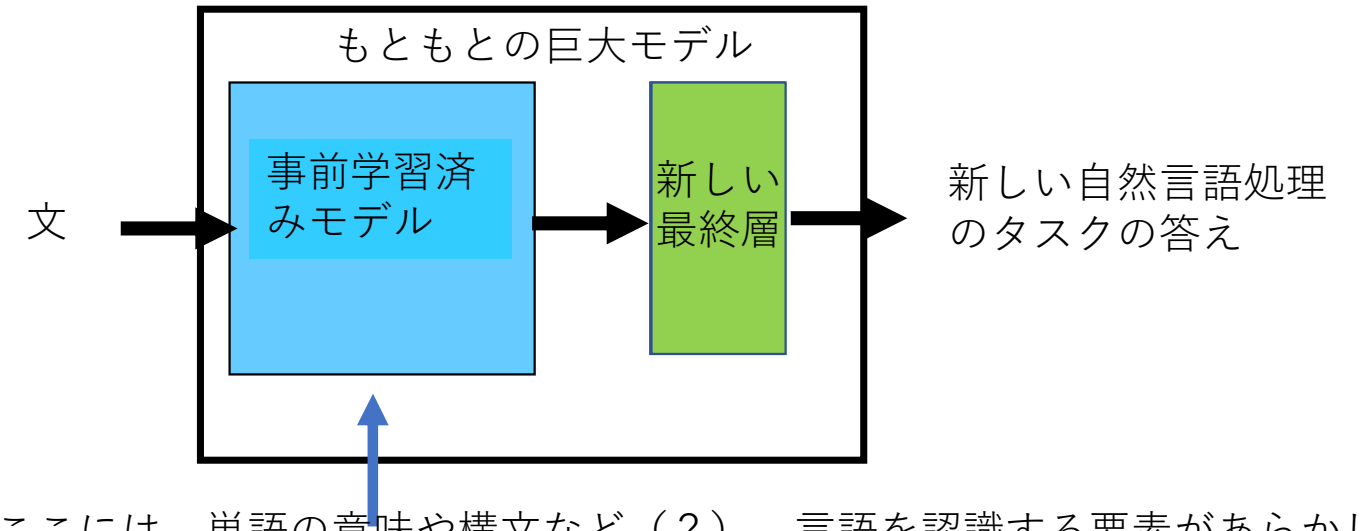
Computer Vision（物体認識の分野）での事前学習済みモデルは前からあった



ここには、斜めの線のパターンなど、画像を認識する要素があらかじめ学習されていて、それは別の物体を認識するときも使えるはず

大規模言語モデル

- Computer Visionにヒントを得て、これを言語でもやる



ここには、単語の意味や構文など（？）、言語を認識する要素があらかじめ学習されていて、それは別の自然言語処理のタスクにも使えるはず

新たなタスクのために（最終層をすげ替えてから）全体の重みを少し変えればいい→この作業を微調整「Fine-tuning」と呼ぶ

自然言語処理の単語の表現の変遷

対象単語のOne-hot ベクトル

分布仮説による周辺単語の 頻度ベクトル

Word2Vec（語彙ごとの分散表現）

ElMoやBERTによる出現レベルの分散表現

大規模言語モデル（＋Fine-tuning）

chatGPT　と　GPT4

- GPT2　GPT3　とモデルサイズは大きくなっていった
- GPT-3.5　（chatGPT）
- モデルサイズはさらに大きく
  社会的にはインターフェースの改善が大きい

2023/3　GPT4が発表される

RLHF：Reinforcement Learning from Human Feedback
　　人間のフィードバックによる強化学習
手法としては深層強化学習

日本語BERTのFine-tuningを利用（頻出語のみ）（R3　多喜さん卒論＋古宮再実験）

- 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

(1)WSDのみ　　(2)WSDと出典の文書分類のマルチタスク学習

国際会議）Kanako Komiya, Nagi Oki and Masayuki Asahara, Word Sense Disambiguation of Corpus of Historical Japanese Using Japanese BERT Trained with Contemporary Texts, PACLIC 2022, (2022, 10, 20).

古宮再実験　結果　(CHJ-WLSP 2019)

Model	Micro Avg.	Macro Avg.
Simple BERT model	**77.50%**	**72.82%**
Multitask learning	77.24%	72.55%
MFS of random sample	73.79%	69.81%
Tanabe (2020)	74.83%	70.80%
MFS of Tanabe (2020)	75.54%	70.00%

- 有意に正解率が上昇

→現代文のBERTにより近代以前の日本語の語義曖昧性解消は性能が上がる（通時適応可能）

- 出典の文書分類とのマルチタスク学習は有効ではなかった

R4時点でのCHJ-WLSP（CHJ-WLSP 2022）

- 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	175,598	1100	説話集
宇治拾遺物語	120,705	1220	説話物語集
十訓抄	90,177	1252	説話集
雑誌　太陽	46,394	1895-1925	雑誌
国定教科書	154,955	1910	国語の教科書

- 出現回数がCHJのみで1000以上の単語だけを対象にしたLexical Sample Task

古宮再実験　結果　（CHJ-WLSP 2022）

Model	Micro Avg.	Macro Avg.
Simple BERT model	84.68%	84.25%
Multitask learning	**85.17%**	**84.45%**
Our MFS	78.29%	78.20%

- CHJ-WLSP 2019とは異なり、マルチタスク学習が有効（有意）
- データ量が増えたおかげ

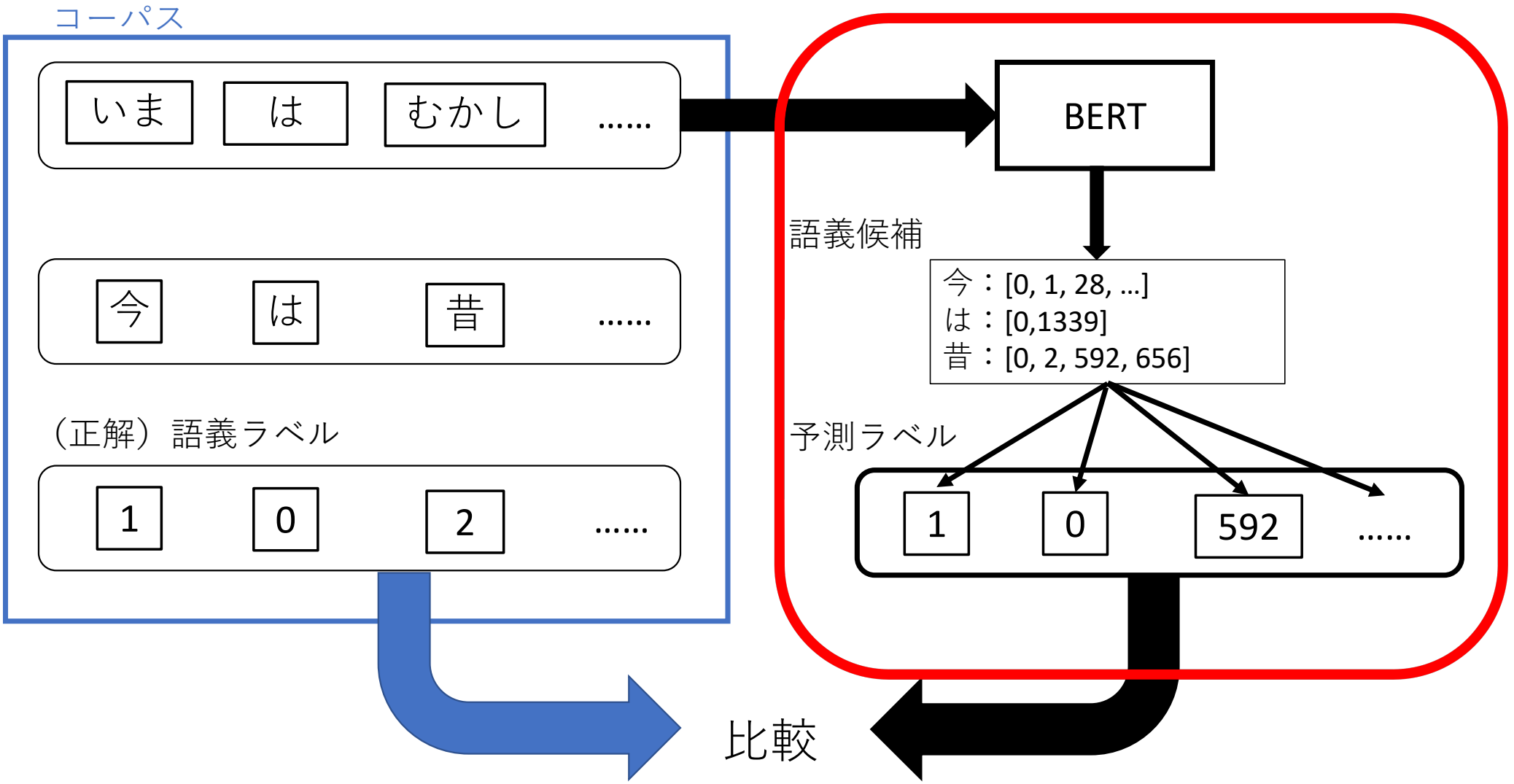
日本語BERTのFine-tuningを利用（全単語）（R4　浅田さん卒論）

- これまでは頻出語のみを取り扱うLexical Sample Task
- 今回は、コーパス中の全単語を扱うall-words WSD

系列ラベリングタスクとして解く

- BERT：東北大のbert-base-japanese-v2　（Unidic区切り）

研究会）浅田宗磨, 古宮嘉那子, 日本語歴史コーパスのAll-words WSD, 言語処理学会第29回年次大会, (2023, 03, 14).
国際会議）Shoma Asada, Kanako Komiya, and Masayuki Asahara (2023), All-Words Word Sense Disambiguation for Historical Japanese, PACLIC 2023, (to appear).


コーパス
いま
は
むかし
……
今
は
昔
……
（正解）語義ラベル
1
0
2
……
BERT
語義候補
今：[0, 1, 28, …]
は：[0,1339]
昔：[0, 2, 592, 656]
予測ラベル
1
0
592
……
比較

実験結果（NLP2023結果）

分類語彙表番号	実験手法	正解率（%）
5桁	MFS	81.61
	提案手法	84.52
3桁	MFS	84.10
	提案手法	87.11

- 開発セットを利用したパラメータチューニングの結果の正解率
- 5桁，3桁のどちらもベースラインとなる最頻出語義（MFS）を3ポイント程上回った
- 国際会議で発表予定の手法ではさらに性能が上がっている

さまざまなコーパスへの語義付与の試み

- 国立国語研究所の浅原正幸教授からの委託研究
- さまざまなコーパスに対してall-words WSDを行う研究を実行中
- 手法：(Asada et al., 2023)　実験は浅田宗磨さんが担当

- 辞書：『分類語彙表増補改訂版』
- 『日本語日常会話コーパス』、『中国語・韓国語母語の日本語学習者縦断発話コーパス』、『多言語母語の日本語学習者横断コーパス』、『名大会話コーパス』、『現日研・職場談話コーパス』、『昭和話し言葉コーパス』、『現代日本語書き言葉均衡コーパス』（BCCWJ）、『昭和・平成書き言葉コーパス』、『日本語歴史コーパス』（CHJ）　＋『日本語話し言葉コーパス』（CSJ）

さまざまなコーパスへの語義付与の試み

- タグ付きデータがあるのはBCCWJとCHJのみ

コーパス	出現数	異なりlemma数	文数
CHJ	647751	19785	24764
BCCWJ	347094	19440	12332

- 現代語コーパスにはBCCWJモデル

BCCWJは87.66%　(5分割交差検定）

- 日本語歴史コーパス（CHJ）には

(1)CHJモデル　(2) BCCWJ＋CHJモデル

CHJは(1)で89.20%　の正解率

語義タグ付きデータにおける間違いの例

（例１）（BCCWJで品詞を間違えている例）

わが国では一昔前までは稀な疾患であったが、**近年**、他のアジア諸国とともに増加傾 向にあり、高齢者の失明疾患として眼科領域で大きな問題となっている。

対象語：近年　予測語義：1.1642　正解語義：3.1642

1.1642の他の語：過去,既往,昔　など

3.1642の他の語：去る,過ぐる,過ぎた　など

（例2）（BCCWJで間違いが頻出した単語の例）

父のようには**なる**まいと、己を律した。

対象語：なる　予測語義：2.1500　正解語義：2.1112

2.1500の他の語：働く,作用する,機能する　など

2.1112の他の語：よる,起因,基因する　など

語義タグ付きデータにおける間違いの例

（例3）（CHJで品詞を間違えている例）

雨は止ず降て**つつ**暗なるに、今一人の男亦着物を脱て裸に成て、前に出でつる男の後に立て出ぬ。

対象語：つつ 予測語義：3.5010　正解語義：1.5010

3.5010の他の語：視覚的,ビジュアル,明るい　など

1.5010の他の語：光,光,光明　　など

（例4）（CHJで間違いが頻出した単語の例）

街談巷説の**中**にも、必ずとるべきことありといへり。

対象語：中 予測語義：1.1720　正解語義：1.1101

1.1720の他の語：範囲,広範囲,裏　など

1.1101の他の語：位,級,等　など

さまざまなコーパスへの語義付与の試み

- その他のコーパスについては、タグ付きデータがないので正解率を求められない
- 現在、一部をサンプリングしてアノテーション中
- 12月には結果が出るので答え合わせを行う

- Asada et al, (2023）の分析によれば、出現回数の少ない単語や、頻出語のうちの出現回数の少ない語義は正解率が低い

→今後の課題

答え合わせ速報

		解析されず	解析誤り	誤り合計	正解率
BCCWJ	図書館書籍	19	15	34	93.2%
BCCWJ	ベストセラー書籍	19	29	48	90.4%
知恵袋		31	26	57	88.6%

500用例ずつ人手チェックした結果