読みコーパスの整備と語義曖昧性解消による読みと意味の頻度表の作成

2025/2/27

東京農工大学　古宮研究室　大井恵奈

目次

- 本研究の背景と目的
- コーパス(文例集)の内容
- それぞれのコーパスへの読みと語義番号の付与方法
- 対応表の作成方法
- 結果と考察

背景

- 近年，音声認識システムや読み上げシステムの需要は高まっている
  - 視覚障がい者の方のための機能
  - 機械音声で作成された動画

- 特に，日本語のシステムでは**漢字**の読みの正確性は大きな課題

背景

例

・　「今日」のおやつは「最中」です.

背景

例

・ 「今日」のおやつは「最中」です.

きょう　　~~さいちゅう~~

こんにち　　もなか

背景

例

・　「今日」のおやつは「最中」です.

・　「きょう」のおやつは「もなか」です.

→話者の**当日のおやつ**が「もなか」である.

・　「こんにち」のおやつは「もなか」です.

→**最近よく食べられているおやつ**は「もなか」である.

どちらの読みも正しいが意味が異なる

関連研究

語義曖昧性解消(Word-sense disambiguation : WSD)

…文中のある単語が，どの語義を表すのかを判定する自然言語処理のタスク

I grow plants.　←植物？工場？

現在は様々な言語で，教師あり学習と大規模な事前学習済みモデルを使用することで一定の性能を達成している．
(Shumidmanら[1]，浅田ら[2])

目的

・『日本経済新聞記事オープンコーパス』，『日本語話し言葉コーパス』を対象に，読みと意味を付与したデータを整備

・日本語の微妙な読みと意味の関係を形態素レベルで明らかにするため，読みと意味の対応表を作成

日本経済新聞記事オープンコーパス[3]

- 2013年1-2月の日本経済新聞朝夕刊 96記事から作成
- 国立国語研究所によりUniDic 形態論情報(短単位・長単位)・文節係り受けを付与
- 加藤らによって語義情報として『分類語彙表』の分類番号が人手で付与されている

短単位形態素数	33,346
長単位形態素数	24,379
文節数	10,627
文数	1,333
記事数	96

日本語話し言葉コーパス (Corpus of Spontaneous Japanese : CSJ)[4]

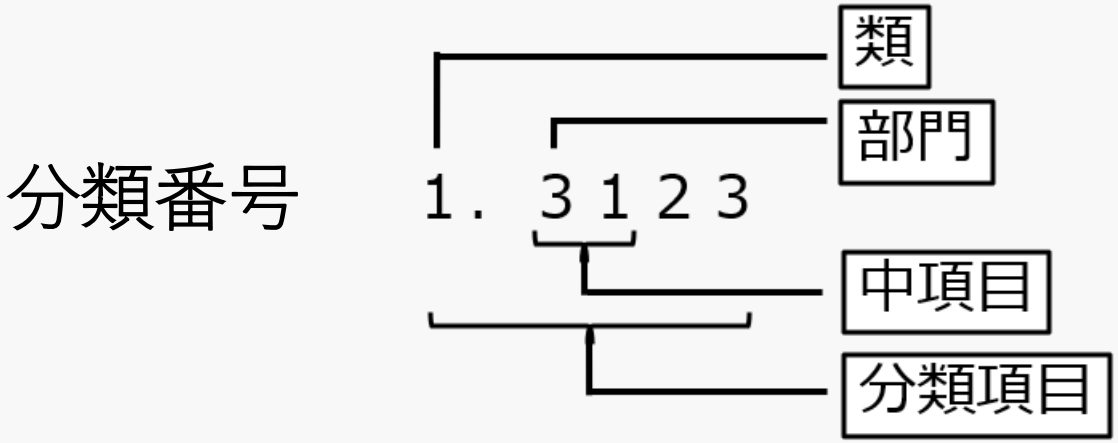
- 日本語の自発音声を大量にあつめて多くの研究用情報を付加した話し言葉研究用のデータベース
- 音声の書きおこしのデータのみで，語義情報が付与されたデータは非公開

転記テキスト	752万語
短単位形態素数	752万
長単位形態素数	631万
文節数	50万

分類語彙表[5]

- 語を意味によって分類・整理したシソーラス (類義語集)

- レコード ID 番号／見出し番号／レコード種別／類／部門／中項目／分類項目／**分類番号**など15項目から成る



本研究では，これを意味として扱う

コーパスと整備方法

- 日本経済新聞記事オープンコーパス

  語義番号は付与されているため，読みを付与する

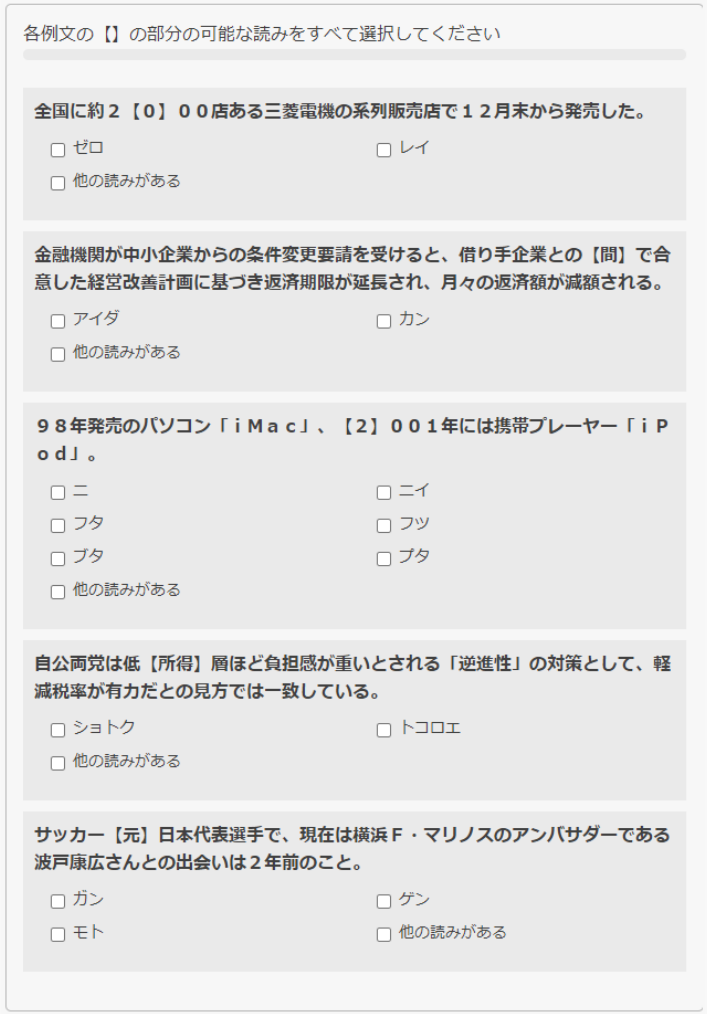
  ＋クラウドソーシング

  ＋専門家によるアノテーション

- 日本語話し言葉コーパス

  音声の書き下しテキストに語義情報を付与する

『日本経済新聞記事オープンコーパス』漢字の読みのクラウドソーシング調査



短単位 33,346語のうち、UniDic で読みの曖昧性がある 4,260語を対象

すべての可能な読みと「他の読みがある」を選択肢として展開

1表現あたり 2人ずつ10回調査

2024/01/01 11:00-13:10 に実施

異なり 1,171 人が参加

2回答あたり2円相当の謝礼

調査概要

- 例文内の単語に対して調査(複数回答可)
  - 例
  ・一→イチ/イッ/ヒト/ビト/他の読みがある
  ・厳しい→イカメシイ/キビシイ/他の読みがある
- 一つの例文に対して複数の単語を対象に読みを質問
  - 直後、再びネイマールが打ったミドルを【止める】と、目を大きく見開いた川島がほえた。
  - 直後、再びネイマールが【打っ】たミドルを止めると、目を大きく見開いた川島がほえた。

問題点と修正

一定数誤った読みを付与する作業者がいる

各解答で外れ値（1件のみ/20件中）を選択した作業者の回答に対して、外れ値を回答するごとに有効票を点数を初期値 1.0 として 0.95 を乗ずる

修正した有効票を割合に変換 (足して 1.0 になるように修正)

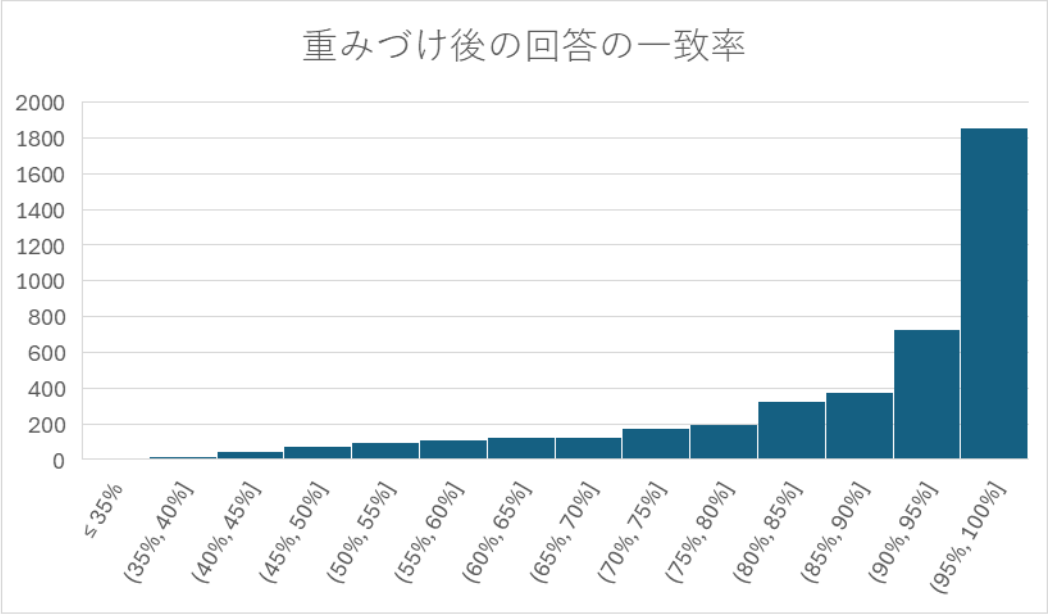
→外れ値の回答が多い回答者の重みが軽くなるような調整

回答が一致したものの件数は448件

回答結果と割合

- 一番多かった回答の割合(重み修正済み)が0.95以上→1857件
  - 横河電、【経常】(ケイジョウ)益４２％増、４～１２月８９億円、資源国向け堅調。…95%
- 一番多かった回答の割合(重み修正済み)が0.90以上→2585件
  - ■石油資源開発　１６日、【米】(ベイ)テキサス州で新型原油「シェールオイル」の鉱区権益を追加取得したと発表した。…90%

今回は，**割合が一番高かった**回答を読みとして採用



『日本経済新聞記事オープンコーパス』漢字の読みの人手によるアノテーション

コーパス内の全ての形態素に対して，Unidicを用いてすべての可能な読みと「他の読みがある」を選択肢として展開し，人手による付与を行う．

※

読みは複数選択可

表層系	読み仮名
その	ソノ
後	アト，ゴ，ノチ
に	ニ
税率	ゼイリツ
が	ガ

表層系	読み仮名
許	キョ，シュー
其亮	キリョウ，チーリャン

短単位では付与できない場合，長単位に付与

表層系	読み仮名
1	センキュウヒャクゴジュウ
9	↓
5	↓
0	↓
年	ネン
大会	タイカイ

略語は一般的な読み仮名を付与

表層系	読み仮名
日本	ニッポン，ニホン
の	ノ
DF	ディフェンダー
を	ヲ
抜き去り	ヌキサリ

『日本語話し言葉コーパス』への語義番号付与

浅田ら[2]の提案手法により，日本語BERTのFine-tuningを利用して，コーパス内の単語全てを対象とするWSD(all-words WSD)をCSJに行った.

CSJ

表音	表層	…	分類番号
マズ	先ず		3.1650
ハッピョウ	発表		1.3140
ハ	は		―
サイショ	最初		1.1651

新たに付与

コーパスと整備方法(再掲)

- 日本経済新聞記事オープンコーパス

  語義番号は付与されているため，読みを付与する

  ＋クラウドソーシング

  ＋専門家によるアノテーション

- 日本話し言葉コーパス

  音声の書き下しテキストに語義情報を付与する

対応表の作成

- 以上の整備したコーパスから，読みと語義の組を抽出
- 複数読みがある場合は，全ての組を作成

その【後】に税率が１０％を超える

【後】……1.1650

作成する読みと語義の組

ゴ―1.1650

アト―1.1650

ノチ―1.1650

- 全ての読みと語義の組を数え，表にする

後	3.1670	1.1650	1.1670
ゴ		+1	
アト		+1	
ノチ		+1	

対応表(例:日本語話し言葉コーパス)

掘る	2.3822
ホル	1

両端	1.1960
リョウタン	4
リョウハシ	1

	読みが一つ	読みが複数
語義が一つ	6102	68
語義が複数	493	70

用法	1.1960	1.3081
ヨウホウ	1	5

略	3.1920	1.1251
リャク	0	1
ホボ	130	0

対応表の件数

日本経済新聞記事オープンコーパス

- クラウドソーシング

	読みが一つ	読みが複数
語義が一つ	375	32
語義が複数	50	21

- 人手によるアノテーション

	読みが一つ	読みが複数
語義が一つ	371	33
語義が複数	50	24

日本語話し言葉コーパスの例

	読みが一つ	読みが複数
語義が一つ	6102	68
語義が複数	493	70

考察－「後」

日本語話し言葉コーパス

後	1.1643	1.1650	1.1670	1.1740	3.1670
コウ	0	0	1	0	0
ゴ	0	9	63	0	0
アト	7	39	297	10	14
ノチ	0	2	28	0	1

分類語彙表(一部抜粋)

分類番号	類・部門・中項目・分類項目	用例
1.1643	体・関係・時間・未来	先，先週，今後，明日
1.1650	体・関係・時間・時間的順序	次，初，順，優先
1.1670	体・関係・時間・時間的	食前, 食後, 前日
1.1740	体・関係・空間・左右・前後・たてよこ	前方, 門前
3.1670	相（形容詞的）・関係・時間・時間的前後	まだ，以後，もう

日本経済新聞記事オープンコーパス

- クラウドソーシング

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	0	2	1
ノチ	1	0	1

- 人手によるアノテーション

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	2	2	1
ノチ	2	0	1

考察－「後」

日本語話し言葉コーパス

左右・前後・たてよこ

未来　時間的順序　時間的　／　時間的前後

後	1.1643	1.1650	1.1670	1.1740	3.1670
コウ	0	0	1	0	0
ゴ	0	9	63	0	0
アト	7	39	297	10	14
ノチ	0	2	28	0	1

分類語彙表(一部抜粋)

分類番号	類・部門・中項目・分類項目	用例
1.1643	体・関係・時間・未来	先，先週，今後，明日
1.1650	体・関係・時間・時間的順序	次，初，順，優先
1.1670	体・関係・時間・時間的	食前, 食後, 前日
1.1740	体・関係・空間・左右・前後・たてよこ	前方, 門前
3.1670	相（形容詞的）・関係・時間・時間的前後	まだ，以後，もう

日本経済新聞記事オープンコーパス

- クラウドソーシング

時間的順序　時間的　時間的前後

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	0	2	1
ノチ	1	0	1

- 人手によるアノテーション

時間的順序　時間的　時間的前後

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	2	2	1
ノチ	2	0	1

考察－「後」
日本経済新聞記事オープンコーパス

- クラウドソーシング

時間的順序　時間的　時間的前後

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	0	2	1
ノチ	1	0	1

- 人手によるアノテーション

時間的順序　時間的　時間的前後

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	2	2	1
ノチ	2	0	1

人手によるアノテーションの方が，複数の読みを回答することが多い

日本経済新聞記事オープンコーパスでは，時間的関係を表す「後」は「ノチ」と読まない

考察－「後」コーパス間の違い

日本経済新聞記事オープンコーパス

クラウドソーシング

時間的順序　時間的　時間的前後

後	1.1650	1.1670	3.1670
ゴ	1	1	1
アト	0	2	1
ノチ	1	0	1

日本語話し言葉コーパス　左右・前後・たてよこ

未来　時間的順序　時間的　　時間的前後

後	1.1643	1.1650	1.1670	1.1740	3.1670
コウ	0	0	1	0	0
ゴ	0	9	63	0	0
アト	7	39	297	10	14
ノチ	0	2	28	0	1

話し言葉コーパス(口語)では時間的関係の「後」を「ノチ」と読むこともある

・書き言葉と話し言葉の違い

・コーパスの大きさの違い

まとめ

- 日本経済新聞記事オープンコーパスには読みをクラウドソーシングと人手によるアノテーションで付与した
- 日本語話し言葉コーパスには分類番号を自動で付与した

→二つのコーパスに語義と読みを整備

- 整備したコーパスから，日本経済新聞記事オープンコーパスからは478件，CSJからは6,733件の語について，読みと意味のペアの出現数の表を作成

- コーパスによって，異なる読みの分布が得られた

→読みの辞書作成や，読み推定システムへの利用