令和 3 年度茨城大学大学院理工学研究科情報工学専攻　修士学位

論文

短単位・長単位の語義を用いた日本語 all-words WSD システム

所属　　情報工学専攻

著者　　20NM730X 溝口賢治

指導教員　　新納浩幸教授

令和 4 年 2 月 4 日 (金)

令和 3 年度茨城大学大学院理工学研究科情報工学専攻　修士学位論文

短単位・長単位の語義を用いた日本語 all-words WSD システム

著者

20NM730X 溝口賢治

指導教員

新納浩幸教授

論文要旨

言葉の意味をその使われている文脈から一意に決定するタスクを語義曖昧性解消（Word sense disambiguation WSD）といい、特にコーパス中の全単語の語義を対象にしたタスクを all-words WSD タスクという。現代日本語書き言葉均衡コーパスでは言葉の単位として、用語収集を目的とした短単位、言語的特徴の解明を目的とした長単位の二種類の単位を用いている。現在の語義曖昧性解消タスクでは、主に短単位を対象とした研究が行われてきた [1]。しかし、長単位の語義が付与されたコーパスによって [2]、長単位の語義の曖昧性を解消するタスクを解くことが可能になった。本研究では、現代日本語書き言葉均衡コーパスにおける長単位を対象とした語義曖昧性解消（all-words WSD）を行う。しかし、長単位はその性質上、語彙数に対する学習データが少ないという課題があるため、学習において短単位のデータを利用することで、長単位を対象とした語義曖昧性解消タスクの正解率の向上を期待する。本研究では LSTM を用いて長単位の語義曖昧性解消を系列ラベリング問題として解き [3]、短単位の語彙の持つ情報を素性として用いた時の長単位の語義曖昧性解消における影響について考察した。

実験では、現代日本語書き言葉均衡コーパスにおいて短単位を単語境界として実験データを作成し、語義曖昧性解消タスクを行った。実験のツールとして、Deep-CRF [4] を利用した。また、基本素性として、短単位語彙素、語種、品詞、品詞細分類、短単位書字形、長単位有無を用いて実験を行った。

実験の結果、Precision は 16.819、Recall は 15.126、F_measure は 15.792 という結果となった。　実験結果から、Deep-CRF の語義予測時に候補となる語義を限定することによる正解率の向上手法を提案した。

Master's Thesis in Scholastic 2021, Major in Computer and Information Sciences, Granduate school of Science and Engineering, Ibaraki University

A Japanese all-words WSD system using short and long unit word senses

Author

20NM730X Kenij Mizoguchi

Advisor

Prof. Hiroyuki Shinnou

Abstract

The task of uniquely determining the meaning of a word based on the context in which it is used is called word sense disambiguation WSD, and in particular, the task for all word senses in the corpus is called the all-words WSD task. The balanced corpus of written Japanese uses two types of units for words: short units for collecting terms and long units for clarifying linguistic features. In the current word sense disambiguation task, research has been conducted mainly on the short unit [1]. However, a corpus with word senses assigned to long units [2] makes it possible to solve the word sense disambiguation task for long units. In this study, we perform word sense disambiguation (all-words WSD) for long units in the balanced corpus of modern Japanese written language. However, due to the nature of long units, there is a problem that the number of training data is small compared to the number of vocabulary. In this study, we used LSTM to solve the word sense disambiguation of long units as a series labeling problem [3], and discussed the effect of using the information of short units as features in the word sense disambiguation of long units.

In this experiment, we created experimental data using short units as word boundaries in a balanced corpus of modern Japanese written language, and conducted a word sense disambiguation task. Deep-CRF [4] was used as a tool for the experiment. We also conducted experiments using the following basic features: short unit lexical feature, word type, part of speech, part-of-speech subdivision, short unit written form, and long unit presence/absence.

The results of the experiment showed that Precision was 16.819, Recall was 15.126, and F_measure was 15.792. Based on the experimental results, we proposed a method to improve the correctness of Deep-CRF word prediction by limiting the number of candidate word senses.

目次

1 序論

1.1 語義曖昧性解消

語義曖昧性解消とは、文章中の多義語の語義を一意に定めるタスクであり、自然言語処理における重要なタスクの一つである。語義曖昧性解消タスクは、頻出単語のみを対象とする Lexical sample task と文中すべての語を対象とする All-Words task の二種類に分類される。Lexical sample task では教師付きデータを用いて各単語タイプごとに分類器を作成するのに対し、All-Words task では単語ごとに十分な訓練データを用意するのが困難なため、すべての語義タグに対して分類器を作成することは現実的ではない。

1.2 関連研究

Luyao Huang ら [5] の研究では、語義の定義文（グロス）を利用することによるニューラルモデルの変化は小さいということを踏まえたうえで、教師あり語義曖昧性解消モデルに効果的にグロスを利用する手法の研究を行った。Marco Maru ら [6] は、人手で曖昧性を解消した語彙的、意味的なリソースである SyntagNet というリソースを提案し、これを用いて教師あり手法による最大の性能を達成した。Christian Hadiwinoto ら [7] は、ElMo や BERT のような文脈をとらえた分散表現の語義曖昧性解消タスクへの効果的な利用手法について、BERT を用いた隠れ層の文脈利用を行った。

1.3 本研究の目的

現代日本語書き言葉均衡コーパスにおいて、言葉の単位は、用語収集を目的とした短単位、言語的特徴の解明を目的とした長単位の二種類の単位が使用されている。現在の語義曖昧性解消タスクでは、短単位を対象とした研究が行われてきた [1]。

本研究では、LSTM を用いて長単位を対象とした語義曖昧性解消を行う。しかし長単位はその性質上、頻出する語が少なく、語彙数に対して学習データが少ないという問題がある。その際に短単位の持つ情報を用いることで長単位の語義曖昧性解消に与える影響を調査した。また、得られた結果から Deep-CRF の単語予測時に候補となる語義を限定する手法による正解率の向上を目的とする。

2　現代日本語書き言葉均衡コーパス

2.1　現代日本語書き言葉均衡コーパスの概要

現代日本語書き言葉均衡コーパスは国立国語研究所コーパス開発センターによって構築されたコーパスである。[*1] 現代日本語書き言葉均衡コーパスは現在、日本語について入手可能な唯一の均衡コーパスとなっている。現代日本語書き言葉均衡コーパスのすべてのサンプルは長単位、短単位の 2 つの言語単位を用いて形態素解析されており、文書構造に関するタグや精密な書誌情報などが提供されている。それぞれの語は No, サブコーパス, サンプル, 開始位置, 終了位置, 短単位語彙素, 語種, 品詞 1, 品詞 2, 短単位書字形, 語彙素番号, 短単位選択, 短単位記入, 短単位分類語彙素番号, 長単位有無, 長単位, 長単位語種, 長単位品詞, 長単位書字形, 長単位選択, 長単位記入, 長単位分類語彙表番号という要素で構成されている。現代日本語書き言葉均衡コーパスにおける語「猫」の構成例を表 1 に示す。

表 1　現代日本語書き言葉均衡コーパスにおける「猫」の例

No	サブコーパス	サンプル	開始位置	終了位置	短単位語彙素
00041_B_PB48_00016	PB	PB48_00016	48780	48790	猫

語種	品詞 1	品詞 2	短単位書字形	語彙素番号	短単位選択
和	名詞-普通名詞-一般	NaN	ネコ	28806	1.5501

短単位記入	短単位分類語彙表番号	長単位有無	長単位	長単位語種
NaN	1.5501	0	猫	和

長単位品詞	長単位書字形	長単位選択	長単位記入	長単位分類語彙表番号
名詞-普通名詞-一般	猫	NaN	NaN	NaN

現代日本語書き言葉均衡コーパスでは用語収集を目的とした短単位、言語的特徴の解明を目的とした長単位の 2 種類の言語単位に分割され、品詞などの情報が付与されている。短単位、長単位のついて、国立国語研究所コーパス開発センターではそれぞれ次のように定義されている。

[*1] https://pj.ninjal.ac.jp/corpus_center/bccwj/

2.2 短単位

現代日本語書き言葉均衡コーパスにおける短単位とは、言語の形態的側面に着目して規定された言語単位である。短単位の認定に当たっては、まず現代語において意味を持つ最小の単位（以下、最小単位）を規定する。そのうえで、最小単位を文節の範囲内で短単位の認定規定に基づいて結合させることにより、短単位の認定を行っている。短単位は基準が分かりやすく, 作業上のゆれが少ないという特徴がある。これは、短単位の基礎となる最小単位の認定において、人によって捉え方の揺れがある要素のない基準を用いているためである。このとき、最もゆれの少ない最小単位をそのまま言語単位としないのは、最小単位では文脈から遠い情報が含まれてしまい、日本語の研究において使いずらい問題があるためである。例として、「気持ち」は「気」と「持ち」の二つの最小単位に分割される。最小単位で解析されたコーパスでは「持つ」の検索結果として「荷物を持つ」といった意味に加え、「気持ち」の意味が得られてしまい、動詞「持つ」の分析を行う際にふさわしくない情報が含まれてしまう。　短単位には、代表形、代表表記、品詞、活用型、活用形が与えられている。短単位への分割及び情報付与には解析用辞書として Unidic が用いられている。[*2]

2.2.1 最小単位規定

最小単位とは現代語において意味を持つ最小の単位のことであり、それぞれ種類ごとに表 2 のように認定され、認定された最小単位は表の 3 ように分類されている。

表 2　最小単位規定

種類	例
和語	/ 豊か / な / 暮らし /
漢語	/ 国 / 語 /
外来語	/ コール / センター /
人名	/ 徳川 / 光圀 /
地名	/ 茨城 / 県 / 水戸 / 市 /
記号	/ 図 / A /

[*2] https://unidic.ninjal.ac.jp/

表 3　最小単位の分類

分類		例
一般		和語: 豊か な 暮らし 漢語: 国 語 外来語: コール センター
数		一 十 百
その他	付属要素	接続的要素: 相 御 各 接尾的要素: 兼ねる がたい 的
	助詞・助動詞	て ます
	人名・地名	徳川 光圀
	記号	A B

2.2.2　短単位規定

短単位の認定規則は表 3 の分類ごとに定められている。認定規定に基づいて最小単位を結合させることにより、短単位の認定が行われている。

短単位認定規定を以下に示す。ここでは、短単位の区切りを「=」、長単位の区切りを「|」で表す。

分類「一般」において、短単位認定規定は以下のように規定される。

- 和語・漢語は 2 最小単位の 1 次結合を 1 短単位とする。
  例: | 母 = 親 |　| 食べ = 歩く |　| 言 = 語 | 資 = 源 |
- 外来語は 1 最小単位を 1 短単位とする。
  例: | コール | センター |　| オレンジ | 色 |

この場合の例外規定は以下のように規定される。

- 省略された外来語の扱い
  - 省略された外来語の最小単位は、和語・漢語の最小単位と同様に扱う。
    例: | パソ = コン |　| 塩 = ビ |　| ピン = ぼけ |
  - 省略された外来語の最小単位と省略されていない外来語の最小単位との 1 次結合体は 1 短単位とする。
    例: | エア = コン |　| マス = コミ |

- 1 最小単位を 1 短単位とするもの
  - 最小単位が 3 個以上並列した場合の各最小単位。

    例: | 衣 | 食 | 住 |　| 松 | 竹 | 梅 |　| 都 | 道 | 府 | 県 |
  - 類概念を表す部分と名を表す部分とが結合してできた固有名詞のうち、類概念を表す部分と名を表す部分とがともに 1 最小単位の場合の、それぞれの最小単位。

    例: | さくら | 屋 |　| 歌舞伎 | 座 |　| のぞみ | 号 |
- 最小単位の 3 個以上の結合体を 1 短単位とするもの。
  - 3 個以上の最小単位からなる組織名などの略称。

    例: | 日経連 |　| 通総研 |
  - 切る位置が明確でないもの、あるいは切った場合とひとまとめにした場合とで意味にずれがあるもの。

    例: | 大統領 |　| 不可解 |　| 明後日 |

    ただし、二つ以上の漢語の最小単位が並列して、1 短単位と結合している場合は、次のように短単位を認定する。

    例: | 中 | 小 | 企業 |　| 都 | 道 | 府 | 県 | 知事 |

分類「数」において、短単位認定規定は以下のように規定される。

- 数以外の最小単位と結合させない。「数」同士の結合は、一・十・百・千のとなえをとる桁ごとに 1 短単位とする。「万」「億」「兆」などの最小単位は、それだけで 1 短単位とする。小数部分は 1 最小単位を 1 短単位とする。

  例: | 十 | 月 | 二十 | 三 | 日 |　| 七百 | 五十 | 二 | 万—語 |

分類「その他」において、短単位認定規定は以下のように規定される。

- 1 最小単位を 1 短単位とする。
  - 付属要素

    例: | 筒 | 状 |　| 扱い | 兼ねる |
  - 助詞・助動詞

    例: | 豊か | な | 暮らし | に | つい | て |
  - 人名

    例: | 徳川 | 光圀 |
  - 地名

    例: | 茨城 | 県 | 水戸 | 市 |

– 記号

例: | 図 |A|

2.3 長単位

現代日本語書き言葉均衡コーパスにおける長単位とは文節をもとにした単位である。長単位の認定には、文節の認定を行ったうえで、各文節の内部を規則に基づいて自立語と付属語に分割していくという手順で行っている。長単位では、複合語を構成要素に分割することなく全体で一つとして扱っている。このような長単位を用いることで、各分野の特徴的な語を把握することができる。

2.3.1 文節規定

文節は一般に付属語または付属語連続の後に境界が存在する。現代日本語書き言葉均衡コーパスではこれに加え、複合辞を付属語として認定している。また、文節の認定の上で問題となる固有名、動植物名、連語の扱いについては、内部の助詞、助動詞で文節を切らないと定められている。

2.3.2 長単位規定

長単位は文節を規定に基づいて分割することによって長単位を認定している。よって長単位が文節を超えることはない。

長単位認定規定を以下に示す。ここでは、短単位の区切りを「=」、長単位の区切りを「|」で表す。

- 記号の規定
  - 区切り符合は 1 長単位とする。

    例: | 湾岸戦争後 |、| 英 |、| 仏 | など | と |
  - 語と同じ働きをする記号・記号連続及びそれらを含む結合体は、全体で 1 長単位とする。

    例: |2000=m2| |WHO|
- 付属語の規定
  - 複合辞を含め付属語は 1 長単位とする。

    例: | 公害紛争処理法 | における | 公害紛争処理 | の | 手続 | は |、| 原則 | として | 紛争当事者 | から | の | 申請 | によって | 開始さ | れる |。|
- サ変動詞の規定
  - 体言及び副詞に形式的な意味の「する」「できる」「なさる」「いたす」が直接続く場合、

体言及び副詞とそれらとを切り離さない。

例: | 往復運動 = し | ている |　 | きちんと = できる |

- 並列の規定
  - 並列の関係にある語は切り離さない。

    例: | 公正妥当 | な | 実務慣行 |
  - 並列の関係にある体言連続のうち、並列された体言全体をうける、またはそれら全体にかかる体言的な形式や接辞がある場合及び形式的な意味の「する」「できる」「なさる」「いたす」がある場合も切らない。

    例: | 英語 = 日本語-間 |　 | 在学 = ・ = 在校する |
- 同格の規定
  - 同格の関係にある体言連続は切らない。

    例: | 機関誌 = 計量国語学 | が | 発刊さ | れ |
- 数量表現の規定
  - 数を表す要素は、単位の変わり目の後ろで切る。

    例: | 平成 |15 年 |9 月 |15 日 | 午後 |7 時 |33 分 |
  - 数を表す要素の前で切る。

    例: | 延べ |23 時間 |30 分 |

2.4　分類語彙表番号

現代日本語書き言葉均衡コーパスには、分類語彙表をもとに短単位・長単位に対して短単位分類語彙表番号・長単位分類語彙表番号が付与されている。分類語彙表における分類番号は「類（1 桁)」.「中項目（2 桁)」「分類項目（2 桁)」からなる数字である。特に 2 桁目を「部門」とよぶ。

「類」は品詞での分類を意味し、分類番号の最初の 1 桁に対応する。「部門」は「類」の下の分類で、意味的に大きなまとまりで分類したものである。「中項目」は「部門」より小さな意味のまとまりで分類したものである。

3　LSTM を用いた語義曖昧性解消

3.1　LSTM の概要

Long Short Time Memory（LSTM）は長期的な依存関係を学習することができるように設計された Recurrent Neural Network の一種である。RNN は従来のニューラルネットワークでは実

現できなかった持続性を持った記憶を保持することが可能であるが、長期系列の学習によって過去データに対する重みが消失してしまうという問題点が存在する。LSTM はこの問題に対する対処として開発された。

LSTM は RNN の中間層のユニットを LSTM Block と呼ばれるメモリと 3 つのゲートを持つブロックに置き換えることで実現される (図 1)。　LSTM Block の構造を図 2 に示す。

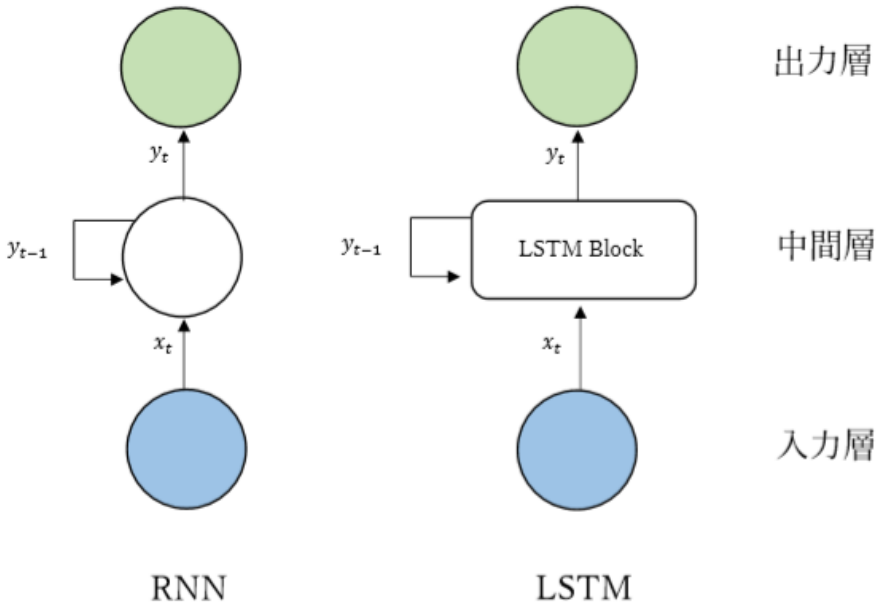

図 1　RNN・LSTM の構造

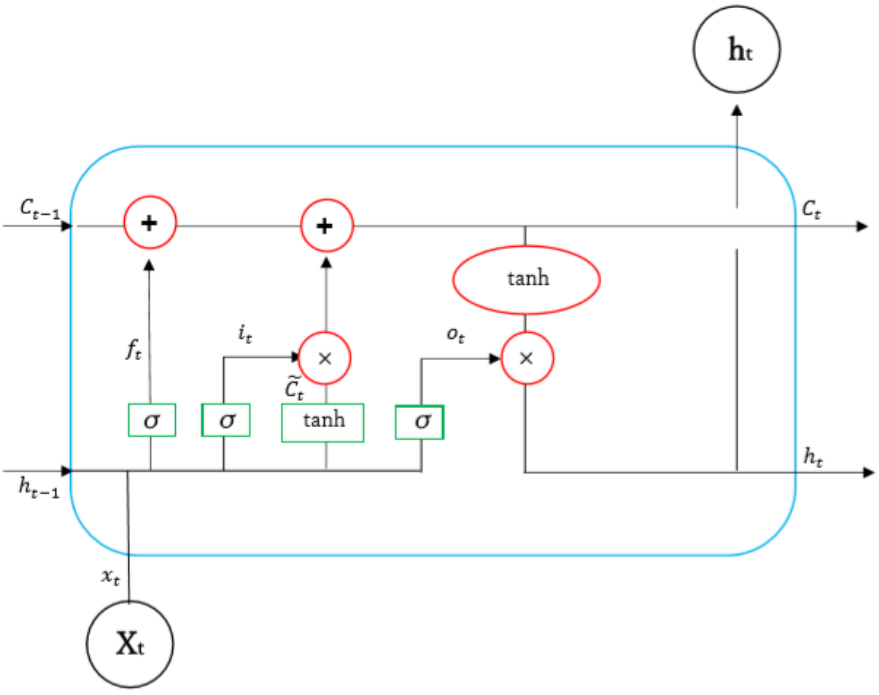

図 2　LSTM Block の構造

3.1.1 LSTM Block

LSTM Block はゲートと呼ばれる情報の取捨選択機構を持つ。各ゲートは以下のような選択機構を持つ。

- Input gate: 前のユニットの入力から受け取るデータの選択
- Forget gate: 記憶しているデータを保持するかの選択
- Output gate: 次のユニットへ出力するデータの選択

Forget gate では、前セルからの長期記憶に対して、情報の取捨選択を行う。f_t はシグモイド関数により 0 から 1 の値が出力される（式 (1)）。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

Input gate では、入力データに対して、どのデータをどの程度の重みで記憶するかを制御する。i_t は前セルから無関係な情報によって誤った重みに更新されることを防ぐため、必要な情報のみが伝搬するように制御する（式 (2)）。また、$\tilde{C}_t$ は入力データから内部で追加される候補値への変換を行う（式 (3)）。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

Output gate では、出力データに対して、どのデータをどの程度の重みで伝搬させるかを制御する。Input gate と同様に、次セルに無関係な情報が伝搬されることを防ぎ、必要な情報のみが伝搬するように制御する。前セルからの長期記憶 C_{t-1} に短期記憶 $\tilde{C}_t$ を加えた C_t（式 (4)）を tanh の入力として用い、h_t（式 (6)）は次セルを活性化させるための重みを決定する。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

3.2 長単位の語義曖昧性解消

本研究では、all-words WSD を各長単位の単語に対して長単位分類語彙表番号という語義を与える系列ラベリング問題とみなし [3]、LSTM を用いてこの系列ラベリング問題を解く。しかし、長単位はその性質上、頻出する語が少なく、語彙数に対して学習データが少ないという問題があるため、短単位の情報を学習に用いて行う。

4　短単位の情報を用いた長単位の語義曖昧性解消

4.1　短単位を単位とした長単位の語義曖昧性解消

長単位データの利用において語彙数に対する学習データが少ないという問題を踏まえ、現代日本語書き言葉均衡コーパスで付与されている短単位のデータを学習に利用することで、長単位語義曖昧性解消タスクの精度向上を期待する。本研究では、語義曖昧性解消の対象を長単位にしぼり、長単位の語義曖昧性解消タスクを解く。このとき、長単位分類語彙表番号の付与されていない短単位の短単位分類語彙表番号の判定を行わないものとした。また、短単位情報を利用するため、入力として短単位区切りのデータを用いた。出力は対象の長単位に対応する長単位分類語彙表番号である。

4.2　実験に用いる基本素性

入力に用いる素性として、現代日本語書き言葉均衡コーパスにおいて付与されている短単位書字形、短単位語彙素、語種、品詞 1、品詞 2、長単位有無を用いて実験を行った。出力は長単位分類語彙表番号である。短単位書字形とは、語が文書中で実際に用いられている形であり、短単位語彙素とは短単位書字形の書字形基本形である。例として、短単位語彙「行く」の短単位書字形として「行く」「行き」「いけ」などがある。語種とは、語を和語・漢語・外来語・混種語・記号に分類したものである。ここで和語とは、日本固有の語を指す。漢語とは、日本語中で音読みで呼ばれる語を指し、音読みで呼ばれる漢字からなる熟語を含む。外来語とは、他言語からなる語を指し、一般にカタカナやアルファベットなどで表記されるものである。混種語とは、異なる複数の語種からなる語である。長単位有無は、その語に長単位が存在する場合に先頭の語に 1、長単位でない語には 0 が付与されている。入出力に用いる基本素性の例「猫」を表 4 に示す。

表 4　入出力に用いる基本素性の例

短単位書字形	短単位語彙素	語種	品詞 1	品詞 2
猫	猫	和	名詞-普通名詞-一般	NaN

長単位有無	長単位分類語彙表番号
0	NaN

5　長単位の語義曖昧性解消の実験

5.1　実験データ

実験データとして、現代日本語書き言葉均衡コーパスから実験データの作成を行った。実験データ作成の手順は以下のとおりである。現代日本語書き言葉均衡コーパスに付与されている短単位書字形、短単位語彙素、語種、品詞 1、品詞 2、長単位有無、長単位分類語彙表番号を取り出し、各素性間の区切りは半角スペースとする。その後、長単位分類語彙表番号以外の素性がデータなしの場合には NULL とする。このとき、長単位分類語彙表番号のないもの（長単位有無が 0 のもの）には O を付与する。文と文の区切りは改行とする。この手順により作成された実験データを longwsd とする。実験データ longwsd は 340880 語、6216 文で構成されている。実験データの構成を表 5 に示す。

表 5　実験データの構成

```
ポケモン ポケモン 外 名詞-普通名詞-一般 NULL 1 B-1.457
カード カード 外 名詞-普通名詞-一般 NULL NULL NULL

「 「 記号 補助記号-括弧開 NULL 0 O
牧師 牧師 漢 名詞-普通名詞-一般 NULL 0 O
を を 和 助詞-格助詞 NULL 0 O
出せ 出す 和 動詞-非自立可能 NULL 0 O
！ ！ 記号 補助記号-句点 NULL 0 O
！ ！ 記号 補助記号-句点 NULL 0 O
」 」 記号 補助記号-括弧閉 NULL 0 O
電話 電話 漢 名詞-普通名詞-サ変可能 NULL 0 O
を を 和 助詞-格助詞 NULL 0 O
取る 取る 和 動詞-一般 NULL 0 O
なり なり 和 助詞-接続助詞 NULL 0 O
罵声 罵声 漢 名詞-普通名詞-一般 NULL 0 O
。 。 記号 補助記号-句点 NULL 0 O
```

5.2　実験設定

実験のツールとして、DeepCRF [4] を用いた。DeepCRF のパラメータとして、以下の数値を与えた。model_name は bilstm-cnn-crf、batchsize は 32、max_iter は 50、n_hidden は 200、

n_word_emb は 100、n_char_emb は 30、n_char_hidden は 30、dropout_rate は 0.33 とした。

本実験では、作成した実験データ longwsd をもとに、文数を基準に 10 分割したデータを作成した。10 分割した実験データを longwsd01、longwsd02、longwsd03、longwsd04、longwsd05、longwsd06、longwsd07、longwsd08、longwsd09、longwsd10 とする。各データの文数は 622 文または、621 文で構成される。各データに対して DeepCRF を用いた実験モデルを作成し、作成した各モデルに対して、評価実験を行った。

本実験では評価指標として、5 分割交差検定を行った。具体的には、各実験データを 5 個のブロックに分割し、4 個のブロックを訓練データとして、1 個のブロックをテストデータとして割り振った。このとき、各ブロックの文数は 125 文または、124 文となっている。また、評価尺度として Precision、Recall、F-measure を用いた。

5.3 評価尺度

本研究では、評価尺度として Precision、Recall、F-measure を用いた。長単位分類語彙表番号について、実験により得られた予測データと正解データの関係は以下のように分類される。

- DeepCRF が付与した長単位分類語彙表番号 (c)
- 正解データに付与された長単位分類語彙表番号 (a)
- DeepCRF が付与した長単位分類語彙表番号のうち、正しく付与されたもの (x)

このとき、Precision、Recall、F-measure をそれぞれ p,r,f とすると、それぞれ式 (7)、式 (8)、式 (9) で与えられる。

$$p = \frac{n(x)}{n(c)} \tag{7}$$

$$r = \frac{n(x)}{n(a)} \tag{8}$$

$$f = \frac{2pr}{p+r} \tag{9}$$

5.4 長単位語義曖昧性解消の実験結果

longwsd01 から longwsd10 までのそれぞれの実験結果について、評価を行った。Precision、Recall、F-measure より、それぞれの値の平均をとり、ALL とする。このとき、Precision は 16.819、Recall は 15.126、F-measure は 15.792 という結果となった。実験結果を表 6 に示す。

表 6　長単位語義曖昧性解消の結果

実験データ	Precision	Recall	F-measure
ALL	16.819	15.126	15.792
longwsd01	14.463	14.085	14.267
longwsd02	9.103	8.938	9.103
longwsd03	23.025	19.612	21.137
longwsd04	17.16	16.875	16.871
longwsd05	24.258	23.529	23.875
longwsd06	14.619	14.252	14.424
longwsd07	10.393	9.447	9.660
longwsd08	5.483	4.415	4.829
longwsd09	18.379	17.754	17.915
longwsd10	29.284	21.168	24.353

6　長単位の語義曖昧性解消の考察

本実験の結果、表 6 より基本素性として現代日本語書き言葉均衡コーパスに付与されている短単位書字形、短単位語彙素、語種、品詞 1、品詞 2、長単位有無を用いた長単位の語義曖昧性解消の結果、Precision、Recall、F-measure は低い数値となった。この要因として、訓練データに含まれていない語（未知語）や訓練データ中の出現頻度の低い長単位分類語彙表番号の存在があげられる。また、分割したデータの内訳によって、精度に大きな差が生じている。このような語に対して、どのように学習、予測を行うべきか考える必要がある。また、DeepCRF が予測した長単位分類語彙表番号が不正解であるときについて、DeepCRF の予測した長単位分類語彙表番号と正解データの長単位分類語彙表番号を比較すると、分類語彙表における「類」や「部門」に相当する桁の数字は一致しているものが大部分を占めていた。よって、不正解となった語に対しても、大まかな予測ができていることが考察される。予測に不正解が存在する例の正解データとの比較を（表 7）に示す。

表 7　予測に不正解が存在する場合の正解データの例

短単位書字形	予測結果の長単位分類語彙表番号	正解データの長単位分類語彙表番号
街	O	O
を	O	O
破壊	B-2.352	B-2.1527
し	NULL	NULL
つくし	NULL	NULL
て	O	O
、	O	O
あたらしく	O	O
魔物	B-1.2301	B-1.203
たち	NULL	NULL
が	O	O
世	O	O

今後の研究として、どの素性を用いたとき長単位語義曖昧性解消に与える影響が大きいかについて調査し、長単位分類語彙表番号の予測に対してその素性を効果的に用いるシステムの構築などが考えられる。また、Fine-Tuninig や短単位・長単位の分類語彙表番号を用いたマルチタスク学習などによる長単位語義曖昧性解消における影響についての調査も必要であると考えられる。本研究では長単位の語義曖昧性解消における短単位情報の利用という実験において、長単位のみを対象として実験を行ったが、追加実験として短単位の語義を対象に含めた実験を行うことで、他システムとの比較を行うことが考えられる。

7　提案手法

長単位の語義曖昧性解消の実験結果より、正解率が低い数値となった要因として DeepCRF が語義予測時にすべての語義から対象となる語の語義予測を行っていることが考えられる。しかし、実際にはある語の候補となる語義は語義ごとに決まっているため、すべての語義から予測する必要はない。そこで、DeepCRF の語義予測時に候補となる語義を限定する手法を提案する。具体的には、現代日本語書き言葉均衡コーパスのもつ長単位の語義を辞書として用いることで、DeepCRF の語義予測時の候補となる語義を限定する過程を追加することでによって、DeepCRF の語義予測の正解率を向上させる。

今後の研究として、DeepCRF に対して提案手法を用いたシステムの構築と追加実験を行い、提案手法の検証や他システムとの比較を行っていく必要がある。

8 BERT を利用した all-words WSD

8.1 BERT

Bidirectional Encoder Representations from Transformers(BERT) とは自然言語処理タスクのネットワークモデルに対する事前学習済みモデルである。大量のラベル無しデータで事前学習を行ったものを少量のラベルありデータでファインチューニングすることで、高精度な結果を出したモデルである。また、多様なタスクに応用でき、少ない学習データで各タスクに対応することが可能であり、文章を双方向に学習することで、「文脈を理解する」能力が高いという特徴がある。

BERT の入出力について、BERT は入力文に対応する単語列を入力とし、埋め込み表現列を出力する。このとき、BERT が出力する埋め込み表現は文脈に依存したものになっている。例として、「私は犬が好きだ。」という文と「お前は警察の犬だ。」という文について考えると、二つの文中の「犬」の語義は異なったものであるが、分散表現データの場合は「犬」の分散表現として同じベクトルを出力する。それに対し BERT の場合は「犬」の周辺単語との関係から文脈に依存した埋め込み表現を出力するため、異なる埋め込み表現となる。

また、BERT はネットワークモデルであるため、対象タスクに応じてネットワークの重みを調整することができ、対象タスクのネットワークの学習時に BERT 自体の学習も同時に行うことが可能である。このとき、対象タスクに応じて調整されることを前提にあらかじめ構築されたモデルを事前学習済みモデルと呼ぶ。このモデルに対して対象タスクに応じたファインチューニングを行うことで事前学習済みモデルを調整する学習が可能である [8]。

8.2 BERT を利用した日本語 all-words WSD システムの構築

BERT を用いた日本語 all-words WSD システムの構築について提案する。

BERT は単語列を入力とし、対応する単語の埋め込み表現を出力する。このとき出力される埋め込み表現は文脈に依存したものになっている。

そこで、BERT から得られる各単語の埋め込み表現を特徴ベクトルとして利用することで、教師あり all-words WSD システムの構築を行う。BERT を利用した all-words WSD のネットワーク図を図 3 に示す。

事前学習済み日本語 BERT モデルとして東北大 bert-japanese を利用し、学習には国立国語研究所の分類語彙表語義付きコーパスを利用する。実験では、文中すべての単語に対し、語義を付与し評価を行う。また、作成したモデルについて評価と本研究を含めた他システムとの比較を行

い、考察される問題点についてシステムの改良を行う。

今後の研究として、BERT を利用した日本語 all-words WSD システムの構築及び評価、実験結果をもとに作成したシステムの改良を行う必要がある。本提案システムと改良したシステムについて及び他システムとの比較を行い、その効果について調査、考察することによって、実用的な日本語 all-words WSD システムの構築を目的とする。

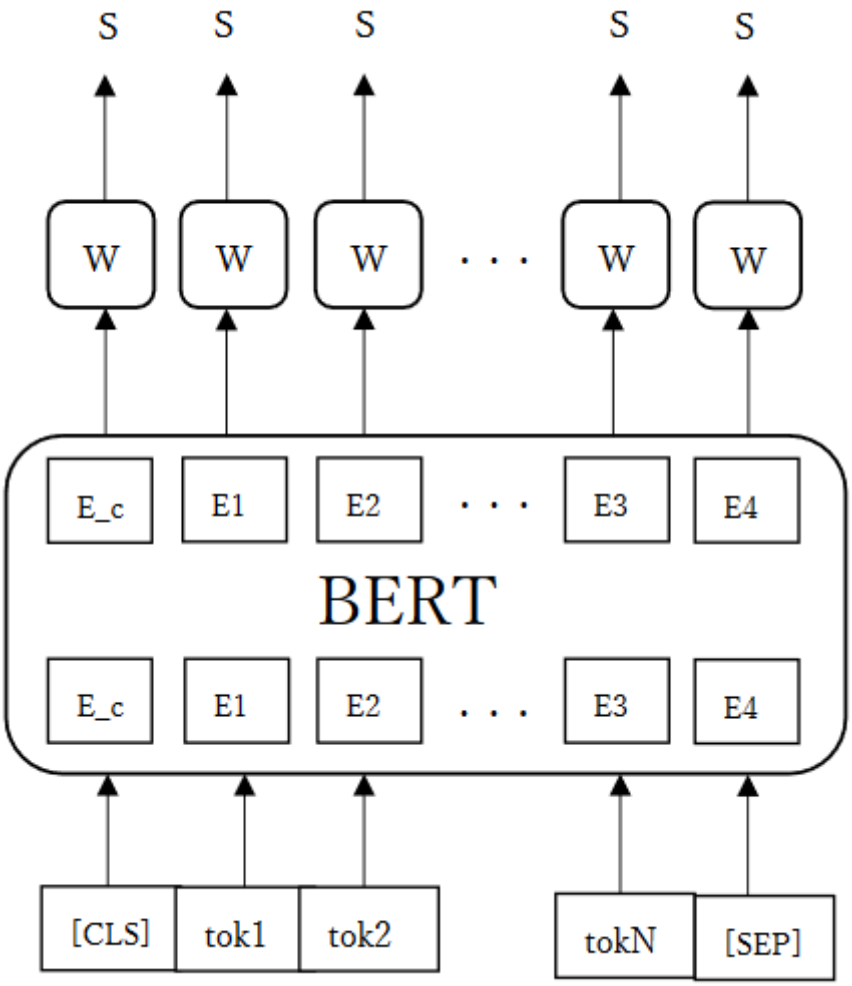

図 3　BERT を利用した WSD のネットワーク図

9　結論

本研究では、LSTM を用いて長単位の語義曖昧性解消を系列ラベリング問題として解き、その際に短単位の持つ情報を素性として用いたときの長単位の語義曖昧性解消における影響の調査を行った。

実験の結果、長単位分類語彙表番号の正解率は低い数値となった。要因ととして未知語や出現頻度の少ない語彙（長単位分類語彙表番号）の存在が考えられ、これらの扱いについてさらなる

工夫が必要になると考察された。また、実験結果から DeepCRF の語義予測時に候補となる語義を限定する手法を提案し、提案手法を用いた追加実験について考察した。さらに、BERT を利用した all-words WSD システムの構築についての提案を行った。

謝辞

担当教員である新納浩幸教授には、本研究への多くの指導や助言をいただきました。この場を借りて御礼申し上げます。

また、本研究にかかわるすべてのお世話になった方々に感謝申し上げます。

参考文献

[1] 鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 概念辞書の類義語と分散表現を利用した教師なし all-words wsd. 自然言語処理, Vol. 26, No. 2, pp. 361–380, 2019.

[2] Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. Annotation of 'word list by semantic principles' labels for the balanced corpus of contemporary written Japanese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December 2018. Association for Computational Linguistics.

[3] 新納浩幸, 鈴木類, 古宮嘉那子. 双方向 LSTM による分類語彙表番号を語義とした all-words WSD. 言語資源活用ワークショップ 2018, 192–202 December 2018.

[4] Motoki Sato. Deepcrf: Neural networks and crfs for sequence labeling, 2018.

[5] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBert: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 3–7 November 2019. Association for Computational Linguistics.

[6] Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. Syntagnet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 3–7 November 2019. Association for Computational Linguistics.

[7] Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 3–7 November 2019. Association for Computational Linguistics.

[8] 新納浩幸. Pytorch 自然言語処理プログラミング. 株式会社インプレス, 2021.