

国立国語研究所学術情報リポジトリ

Analysis of Use of Monosemous Words for Word Sense Disambiguation

メタデータ	言語: jpn 出版者: 公開日: 2021-03-05 キーワード (Ja): キーワード (En): 作成者: 佐々木, 稔, 谷田部, 梨恵, Yatabe, Rie メールアドレス: 所属:
URL	https://doi.org/10.15084/00003162

語義曖昧性解消における辞書に定義された 単義語利用についての分析

佐々木 稔 (茨城大学工学部情報工学科) †
谷田部 梨恵 (茨城大学大学院理工学研究科)

Analysis of Use of Monosemous Words for Word Sense Disambiguation

Minoru Sasaki (Ibaraki University)

Rie Yatabe (Ibaraki University)

要旨

多義語の語義曖昧性解消を自動的に行う際、一般的に周辺の共起単語を特徴として利用する。周辺の文脈情報は語義曖昧性解消を行う際の大きな手掛かりとなるが、周辺文脈には複数の語義を持つ多義語が多く存在する。そのため、多義語により文脈を誤って捉えてしまい、語義曖昧性解消の精度に悪影響を及ぼす可能性がある。そこで本研究では、語義曖昧性解消システムにおいて、辞書に定義された単義語を効果的に使用するための方法について調査し、従来の基本素性ベクトルとどのように組み合わせると有効なのか分析を行う。単義語は語義をひとつしか持たず、語義の特徴が一意に決定されると考えられる。この性質を語義曖昧性解消に応用することで、対象単語の周辺文脈を適切に捉えることができると考えられる。単義語の分散表現、単義語のフラグについて様々な素性の組み合わせについて語義曖昧性解消実験を行い、どの組み合わせが有効なのか分析する。

1. はじめに

本稿では、語義曖昧性解消において、辞書で単義語として定義された単語を特徴とすることの有効性について分析を行う。

語義曖昧性解消は文中の多義語が辞書中のどの語義で使われるのかを識別するタスクである。計算機で語義曖昧性解消を行うには一般的に教師あり学習手法が用いられる。対象単語の語義が付与された訓練データから分類モデルを学習し、得られた分類モデルに語義推定したい用例文を入力し、最もふさわしい語義を選択する。多くの語義曖昧性解消システムでは、対象単語周辺(2~3単語)に出現する単語が分類モデルを学習するための素性として用いられる(Zhong 2010) (新納, 村田, 白井, 福本, 藤田, 佐々木, 古宮, 乾 2015)。周辺の文脈情報は語義曖昧性解消を行う際の大きな手掛かりとなるが、周辺文脈に出現する単語は複数の語義を持つ多義語が多く存在する。そのため、多義語により文脈を誤って捉えてしまうため、語義曖昧性解消の精度に悪影響を及ぼす可能性がある。前後の単語数を替えるとといった素性の拡充を図る工夫も考えられるがすべての単語について有効とは限らず、訓練データに含める用例文の拡充を図る場合には専門家による語義ラベル付与の作業に大きなコストがかかるという課題が存在する。

そこで、本研究では語義曖昧性解消システムにおいて、辞書に定義された単義語を効果的に使用するための方法について調査し、従来の基本素性ベクトルとどのように組み合わせると有効なのかについて分析を行う。単語が単義語かどうかについては岩波国語辞典において語義が1つだけ存在する見出し語を単義語と判断して利用する。また、単義語を効果的に利用するために、ひらがな1文字だけの単義語を取り除いた場合の単義語リストも用意する。これら2種類の単義語リストと様々な素性を組み合わせて、語義曖昧性解消実験

† minoru.sasaki.01@vc.ibaraki.ac.jp

を行い、素性として単義語を利用することが有効かどうかについて分析を行う。

2. 関連研究

辞書などの外部知識から得られる単義語情報を利用した知識ベースに基づく手法もこれまでに研究が行われている。Liらは単義語を語義曖昧性解消の周辺文脈素性として利用する手法を提案し、その有効性を検証している(Li 1999)。しかし、この手法は指定された単義語しか素性として使えないため、利用可能な単語は限定されてしまう。また、ニューラルネットワークに基づいた単語の分散表現を利用しておらず、日本語の用例文における単義語の有効性については検証されていない。

(Leacock et al. 1998)や(Mihalcea and Moldovan 1999)は対象単語の各語義の特徴を捉えるために、各語義について単義の類義語を利用する手法を提案している。まず、対象単語に対する WordNet に登録された各語義について、類義語の集合である synset に含まれる単義語を抽出する。抽出された単義語の定義文から語義の説明部分を抽出し、この部分を検索フレーズとして検索エンジンを用いて用例文の収集を行う。この手法は各語義の特徴として単義の関連語が有効であることを示している。しかし、WordNet のような類義語が登録された辞書を用意しなければならない、単義語が類義語に存在しない語義は使えないといった制約がある。

3. 単義語を用いた語義曖昧性解消システム

本節では、単義語を用いた語義曖昧性解消システムについて述べる。このシステムは対象となる多義語を含む用例文に対して、辞書中の語義の中で最もふさわしい語義の出力を行う。

<p>Headword た 多義語 30471-0-0-1-0 <1>述べている事に関し、話し手・書き手の確認の気持を表すのに使う。 (例)「見つけたぞ」、「さあ、そこをどいた」、「望みは捨てた」 30471-0-0-2-0 <2>動作・作用やその結果が、ある状態として存続する意を表す。 (例)「さびたナイフ」 30471-0-0-3-0 <3>着目言及時より前、つまり過去に属する(そしてそれを経験している)という意を表す。 (例)「そんな話も出た」</p>
--

<p>Headword た 単義語 30473-0-0-0-0《動詞・形容詞の上に》語調を強める。 (例)「たばかり」、「たやすい」</p>
--

図 1: ひらがなの単義語

3.1 単義語の定義

単義語は辞書において語義をひとつだけ持つ単語と定義し、ふたつ以上の語義が列举された単語は多義語として扱う。単義語と多義語を分けるために辞書として岩波国語辞典 ¥cite{iwanami94} を利用する。この辞書は SemEval-2010: Japanese WSD タスクで配布されたデータで、辞書に記載されたすべての語義に語義 ID が付与されている。本研究では、見出し語に対して大分類以下の分類番号がすべて 0([見出し語 ID]-[複合語 ID]-0-0-0)である語義のみをもつ単語を単義語、それ以外の単語を多義語と定義する。

しかし、辞書中に記載されているひらがなの単義語は、語義曖昧性解消の精度に悪影響を及ぼす恐れがある。例えば、対象単語「会う」の用例文「初めて会った日のことを覚えている？」に対して、前後 2 単語以内に出現する単語「初めて」、「た」、「日」が抽出される。このとき、図 1 に示すように、「た」という単語には、辞書において単義語と多義語の両方の語義が登録されている(図 1)。本来、「たやすい」の「た」や「か細い」の「か」が

単義語として判断することが理想である。しかし、辞書において品詞が登録されていないため、助詞の「た」や「か」も単義語として判定されて分類精度が下がってしまうという問題がある。この問題に対処するため、ひらがな 1 文字の単語は単義語として扱わず、残りの単語を単義語として扱うこととする。

ひらがな 1 文字の単義語を取り除く効果を調べるため、辞書で定義された単義語を使った単義語リストを使った場合をAグループ、ひらがな 1 文字を除外した場合をBグループとして各語義曖昧性解消システムに適用する。

3.2 素性

今回は2種類の単義語素性について有効性を分析するために、以下に示す 5 種類(①~⑤)の素性を使用する。

3.2.1 ①ベースライン素性

対象単語を含む用例文に対し、UniDic を辞書として形態素解析を行い、対象単語及び前後二単語の単語、品詞、品詞大分類、係り受け、シソーラス情報を素性として抽出する。これらの素性の詳細を以下の 20 種類(e1~e20)の素性として定義する。

- e1=二つ前の単語, e2=二つ前の品詞, e3=その品詞大分類,
- e4=一つ前の単語, e5=一つ前の品詞, e6=その品詞大分類,
- e7=対象の単語, e8=対象単語の品詞, e9=その品詞大分類,
- e10=一つ後の単語, e11=一つ後の品詞, e12=その品詞大分類,
- e13=二つ後の単語, e14=二つ後の品詞, e15=その品詞大分類,
- e16=係り受け, e17=二つ前の分類語彙表の値, e18=一つ前の分類語彙表の値,
- e19=一つ後の分類語彙表の値, e20=二つ後の分類語彙表の値

分類語彙表の ID は 5 桁のものを使用している。また、一つの単語に対して分類語彙表 ID は複数存在するため、e17~e20 に対する素性は複数存在する。これらの素性に対して、各素性の出現頻度をベクトルの要素として割り当てることで、用例文をベクトル化する。

3.2.2 ②単義語フラグ

3.1 節に示した単義語の定義に従って、出現単語に対して単義語フラグを割り当てる。前後 2 単語以内に出現する各単語に対して、単語が存在すれば単義語であるかどうかを判定し、単義語と多義語の区別を行う。この区別に応じて単義語フラグとして 2 次元ベクトルを定義し、単義語であれば(1,0)、多義語であれば(0,1)と設定する。前後の位置に単語が出現しない場合は(0,0)を割り当てる。

3.2.3 ③対象単語を除く単語の分散表現

分散表現はニューラルネットワークを用いて単語を高次元の実数ベクトルで表現したものである。単語の分散表現の作成には word2vec や GloVe などを使用し、大量のテキストを入力することで、あらかじめ設定した次元のベクトルがすべての出現単語に対して作成される。単語をベクトル化することで、「王様」-「男」+「女」=「女王」や「パリ」-「フランス」+「日本」=「東京」といった意味的な演算処理ができるようになり、単語間の意味に基づいて関係性を捉えることができる。本研究では、国立国語研究所が公開する国語研日本語ウェブコーパスから得られた分散表現である nwjc2vec を用いる^{Asahara18}。nwjc2vec はウェブを母集団とした約 258 億語からなる日本語コーパスに対して word2vec の CBOW モデルを用いて分散表現を求めており、各単語は 200 次元ベクトルで表現されている。

対象単語の前後 2 単語以内に出現する対象単語を除く各単語に対して、出現単語に対して nwjc2vec に分散表現が定義されていれば、その 200 次元ベクトルの分散表現を用いる。

出現単語が存在しない、もしくは単語は存在するが分散表現が定義されていない場合は 200 次元のゼロベクトルを用いる。このように 4 単語の分散表現を素性として利用する。

3.2.4 ④対象単語を含む単語の分散表現

対象単語の前後 2 単語以内に出現する対象単語を含む各単語に対して、出現単語に対して `nwjc2vec` に分散表現が定義されていれば、その 200 次元ベクトルの分散表現を用いる。出現単語が存在しない、もしくは単語は存在するが分散表現が定義されていない場合は 200 次元のゼロベクトルを用いる。このように 5 単語の分散表現を素性として利用する。

3.2.5 ⑤単義語の one-hot ベクトル

単語の分散表現を利用した場合と比較するため、単語の one-hot ベクトルを利用した場合についても考える。対象単語の前後 2 単語以内に出現する単語の総数を次元とするベクトルを設定する。前後 2 単語以内のそれぞれの単語について、その単語に対応するベクトルの要素を 1、その他の要素をすべて 0 とする one-hot ベクトルを作成する。前後 2 単語以内に単語が出現しないときは、すべての要素を 0 とするゼロベクトルとする。

3.3 分類モデル

最もふさわしい語義を推定するための分類モデルとして、多層パーセプトロン(MLP)とサポートベクターマシン(SVM)の 2 種類の機械学習手法を利用する。

3.3.1 ⑥MLP (Multi-Layer Perceptron)

MLP は教師あり学習による多クラス分類モデルとしてよく使われる手法で、入力層・中間層・出力層の 3 つの層から成る階層型ネットワークである。入力層に入力データから得られた多次元ベクトルを渡し、重み付けによって中間層、出力層へと伝わり、出力層の各ノードの値をソフトマックス関数で対応する語義ラベルに分類される確率を出力する。この多層パーセプトロンに教師データを入力し、教師データの語義と学習結果の誤差が小さくなるように、出力層から入力層に向けて重みの修正をすることによって学習が行われる。

今回の実験では中間層のノード数を 100、最適化手法として確率的勾配降下法を使用、学習の繰り返し数であるエポック数を 10,000 と設定して分類モデルの学習を行う。また、MLP は初期値の影響が大きく、実験を行う度に精度が変化するため、同じ実験を 3 回繰り返し、それぞれの平均精度をさらに平均した値を実験結果として使用する。

3.3.2 ⑦SVM (Support Vector Machine)

SVM は二値分類を行う機械学習手法として広く使われている。この SVM は特徴空間において、訓練データ中のあるラベルのデータとその他のラベルのデータの間の距離が最大となるように分離超平面的学習を行う。未知のデータを分類する際は学習で得られた分離超平面を境界としてラベルの判定を行う。

今回の実験では、Python のライブラリである `scikit-learn` の `LinearSVC` を用いて線形の境界面を学習する。学習時の誤差を最小にするための繰り返しの最大回数はデフォルト値の 1,000 では少なかったことから、100 万と設定し、残りのパラメータはデフォルト値で学習を行う。

3.4 分析に使用する語義曖昧性解消システム

3.1 節に示した単義語リストの A グループと B グループの違いで語義曖昧性解消精度がどの程度変わるのかを調べるために、2 つのグループと 3.2 節に示した 5 種類の素性、3.3 節に示した 2 種類の分類モデルを組み合わせることにより、表 1 に示す 17 種類の語義曖昧性解消システムを作成する。

表 1：各語義曖昧性解消システムで使用する素性と単義語，モデル（空欄は不使用）

システム ID	ベースライン素性	単義語フラグ	分散表現, one-hot ベクトル	分類モデル
1	①			⑥
2	①	②A		⑥
3	①	②B		⑥
4	①	②A	③A	⑥
5	①	②B	③B	⑥
6	①		③A	⑥
7	①		③B	⑥
8	①	②A	⑤A	⑥
9	①	②B	⑤B	⑥
10	①	②A	③A	⑦
11	①	②B	③B	⑦
12		②A	③A	⑥
13		②B	③B	⑥
14		②A	④A	⑥
15		②B	④B	⑥
16		②A	④A	⑦
17		②B	④B	⑦

4. 実験

本節では，3.4 節に示した 17 種類の語義曖昧性解消システムの分類精度を分析するために，評価データによる精度比較実験を行う。

4.1 実験データ

本研究における対象単語は，SemEval2010 日本語 WSD タスクデータである対象単語の 50 個を利用する(Okumura, Shirai, Komiya, Yokono, 2010)。また，訓練データとテストデータはその単語を使用した用例文データを現代日本語書き言葉均衡コーパス(BCCWJ)からそれぞれ 50 個用意されている。

5. 実験結果

17 種類の語義曖昧性解消システムに対して，SemEval2010 日本語 WSD タスクデータで実験を行ったときの平均精度を表 2 に示す。表 2 の結果において，ベースライン素性のみを素性とするシステム 1 を用いたときの平均精度は 76.48% で，この素性に分散表現や単義語フラグを追加した場合の精度と比較することにより，単義語の有効性を分析する。最も高い分類精度となったのはシステム 9 の基本素性にひらがな 1 文字の単義語を除いた単義語フラグを用いて単義語フラグ，単義語の one-hot ベクトルを追加した場合であった。この結果は単義語を使わないシステム 0 と有意差検定を行った結果，有意水準 5% で有意な差があった。ベースライン素性に分散表現を追加したシステムはベースライン素性のみを用いたシステムと比べて全体的に精度が低い結果となった。

単義語を用いた場合は前後 2 単語以内に単義語が存在することが少ないため，平均精度が少し変化する結果となった。単義語リストにおいて，ひらがな 1 文字の単義語を含めた場合と除外した場合を比較すると，システム 3 とシステム 13 は精度が低い結果となったが，

それ以外ではひらがな1文字の単義語を除外した方が高い精度が得られた。

表 2: 各システムに対する語義曖昧性解消の平均精度
(ひらがな1文字の単義語を除いた単義語リストによる平均精度は太字で表す)

システム ID	50 単語の平均精度
1	0.7648
2	0.7660
3	0.7652
4	0.7288
5	0.7433
6	0.7195
7	0.7367
8	0.7672
9	0.7705
10	0.7372
11	0.7648
12	0.7493
13	0.7483
14	0.7535
15	0.7611
16	0.7688
17	0.7696

6. 考察

6.1 ひらがな1文字の単義語を含めた場合

表 2 に示すシステム 2, 4, 6, 10, 12, 14 の平均分類精度を見ると、単義語リストにひらがなの単義語を含めた場合は語義の分類精度が著しく低下した。ひらがなの単語は対象単語と共起することが多く、他の見出し語でも使われることから、前後の内容に揺らぎが生じたと考えられる。共起しやすいひらがな単語は 3.1 節で述べた「た」や「か」があり、複数の見出し語を持つ多義語とみなすことができる。例えば、「た」や「か」は助詞としての意味の他に、「たやすい」、「か弱い」といった形容詞の上に付いて語調を整え勢いを強める単義語としての意味を持つ。そのため、対象単語が「一 (イチ)」で「一か八かの勝負」という要例文が入力されると、周辺単語として「か」と「八」が抽出される。この場合の「か」は助詞の単語で複数の語義を持つ多義語である。したがって、単義語の分散表現を用いた語義曖昧性解消を行う際、多義語の単語を誤って単義語として扱ってしまったことから、正しい分類が行われず、分類精度が著しく低下したと考えられる。

6.2 ひらがな1文字の単義語を除外した場合

ひらがな1文字の単義語を除外したシステムは、多くの素性の組み合わせでひらがな1文字を含めた場合よりも平均分類精度が高くなる傾向が見られた。対象単語の前後に出現する単義語を手がかりとして使うことによって、僅かではあるが語義の識別に効果があると考えられる。例えば、岩波国語辞典に記載されている単義語の中に「悪い」という単語がある。対象単語「相手」における「相性が悪い相手」という用例文に対し、対象単語の1単語後に単義語「悪い」が出現する。「相手」と共起する「悪い」という単義語の影響により、「相手」の語義が「物事をするとき、行為の対象となる人。」を選びやすくなる。

6.3 ベースライン素性に単語の分散表現を追加する場合

ベースライン素性に分散表現を追加したシステム(システム 4, 5, 6, 7, 10, 11)はベースラ

イン素性のみを用いたシステムと比べて全体的に精度が低い結果であった。ベースライン素性と分散表現を組み合わせることで、分散表現の密な値がノイズとなって大幅な精度の低下が起こっていると考えられる。そのため、システム 12~17 のようにベースライン素性を使わずに分散表現だけを素性として利用することで、システム 16 や 17 のように高い平均分類精度を出すことが可能だと考える。

7. 結論

本稿では、語義曖昧性解消システムにおいて、辞書に定義された単義語を効果的に使用するための方法について調査し、従来の基本素性ベクトルとどのように組み合わせると有効なのかについて分析を行った。実験の結果、辞書に定義される単義語からひらがな 1 文字の単義語を除外することで分類精度を 1%前後向上することができた。対象単語の前後に出現する単義語を手がかりとして使うことによって、語義の識別に効果があることを示した。また、素性の組み合わせにより、単義語の効果が下がる場合もあることが明らかになった。

今後は、今回扱った以外の素性の組み合わせについても実験を行ってさらなる分析を行うことや、単義語のさらに効果的な利用方法について検討することが課題である。

文 献

- Z. Zhong and H. T. Ng (2010). “It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text”, Proceedings of the ACL 2010 System Demonstrations, pp.78–83.
- 新納浩幸, 村田真樹, 白井清昭, 福本文代, 藤田早苗, 佐々木稔, 古宮嘉那子, 乾孝司 (2015). 「クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け」自然言語処理, 22 (5), pp.319-362.
- J. Li and C. Huang (1999). “A model for word sense disambiguation,” in International Journal of Computational Linguistics & Chinese Language Processing, Volume 4, Number 2, pp. 1–20.
- C. Leacock, M. Chodorow, and G. A. Miller (1998). “Using corpus statistics and WordNet relations for sense identification”, In Computational Linguistics, volume 24, pages 147–165.
- R. Mihalcea and D. Moldvan (1999). “An Automatic Method for Generating Sense Tagged Corpora”, Proceedings of the American Association for Artificial Intelligence (AAAI), 461–466.

関連 URL

『現代日本語書き言葉均衡コーパス』

http://pj.ninjal.ac.jp/corpus_center/bccwj/