平成２７年度茨城大学工学部情報工学科
卒業研究論文

類義語を利用した単語の分散表現から語義の分散表現の構築

平成28年2月9日提出

茨城大学 工学部情報工学科
11t4010l 大内克之
指導教員：新納 浩幸 教授

平成 27 年度茨城大学工学部情報工学科卒業研究

類義語を利用した単語の分散表現から語義の分散表現の構築

氏名：11t4010l 大内 克之
指導教員：新納 浩幸 教授

論文要旨

本論文では語義の分散表現の構築方法を提案する。

単語の分散表現とは、その単語の意味を低次元の密なベクトルで表現したものである。従来の bag of words による高次元の疎なベクトルで表現するよりも、よりよく意味を表現できていると考えられる。そのため様々な自然言語処理のタスクに利用され、多くの成果を出している。

語義曖昧性解消のタスクに対しては、通常、教師付き学習手法が用いられる。しかし教師付き学習手法の場合、訓練データの作成コストが高いことから対象とする単語が限定されてしまい、実用的ではないという問題がある。一方、語義の分散表現を求めることができれば、対象単語の文脈のベクトルとどの語義の分散表現が類似しているかを調べることで語義曖昧性解消が実現できる。単語の分散表現はタグなしコーパスから構築できるため、語義の分散表現も同様の手法から構築できれば教師なしの語義曖昧性解消が実現できることになる。このような背景から語義曖昧性解消に関しては語義の分散表現を構築する試みがなされている。ここでは語義の分散表現を構築するために、多義語の各語義の分散表現の和が、多義語の分散表現になっていると考える。つまり多義語の分散表現を $\boldsymbol{v}$ とし、その多義語の各語義 $s_i(i = 1 \sim K)$ の分散表現を $\boldsymbol{v}_i$ とするとき、

$$\boldsymbol{v} = \sum_{i=1}^{K} \boldsymbol{v}_i$$

が成立していると考える。そして本論文ではこの関係式を利用して $\boldsymbol{v}$ から $\boldsymbol{v}_i$ を構築する方法を提案する。具体的には s_i の語義を持つ類義語 w_i の分散表現 $\boldsymbol{u}_i$ を利用する。$\boldsymbol{u}_i \approx \boldsymbol{v}_i$ と考えられるため $\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$ とし、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

から最小二乗法により α_i を求めることで $\boldsymbol{v}_i$ を構築する。

実験では BCCWJ コーパスから分散表現を求め、単語「意味」が持つ 3 つの語義の分散表現を構築した。この構築した語義を利用して SemEval-2 の日本語辞書タスクにおける「意味」のテスト用例の語義曖昧性解消を行い、59.0% の正解率を得た。

目 次

第1章　序論

1.1　概要

本論文では語義の分散表現の構築方法を提案する。

単語の分散表現とは、その単語の意味を低次元の密なベクトルで表現したものである。従来の bag of words による高次元の疎なベクトルで表現するよりも、よりよく意味を表現できていると考えられる。そのため様々な自然言語処理のタスクに利用され、多くの成果を出している。

語義曖昧性解消のタスクに対しては、通常、教師あり学習手法が用いられる。しかし教師あり学習手法の場合、訓練データの作成コストが高いことから対象とする単語が限定されてしまい、実用的ではないという問題がある。一方、語義の分散表現を求めることができれば、対象単語の文脈のベクトルとどの語義の分散表現が類似しているかを調べることで語義曖昧性解消が実現できる。単語の分散表現はタグなしコーパスから構築できるため、語義の分散表現も同様の手法から構築できれば教師なしの語義曖昧性解消が実現できることになる。このような背景から語義曖昧性解消に関しては語義の分散表現を構築する試みがなされている [3][1]。ここでは語義の分散表現を構築するために、多義語の各語義の分散表現の和が、多義語の分散表現になっていると考える。つまり多義語の分散表現を $\boldsymbol{v}$ とし、その多義語の各語義 $s_i(i = 1 \sim K)$ の分散表現を $\boldsymbol{v}_i$ とするとき、

$$\boldsymbol{v} = \sum_{i=1}^{K} \boldsymbol{v}_i$$

が成立していると考える。そして本論文ではこの関係式を利用して $\boldsymbol{v}$ から $\boldsymbol{v}_i$ を構築する方法を提案する。具体的には s_i の語義を持つ類義語 w_i の分散表現 $\boldsymbol{u}_i$ を利用する。$\boldsymbol{u}_i \approx \boldsymbol{v}_i$ と考えられるため $\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$ とし、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

から最小二乗法により α_i を求めることで $\boldsymbol{v}_i$ を構築する。

実験では BCCWJ コーパス [2] から分散表現を求め、単語「意味」が持つ 3 つの語義の分散表現を構築した。この構築した語義を利用して SemEval-2 の日本語辞書タスク [4] における「意味」のテスト用例の語義曖昧性解消を行い、59.0% の正解率を得た。

1.2 本論文の構成

本論文では提案手法を説明する前に、予備知識として必要なトピックスを扱う。具体的には、多義語とその曖昧性に関する問題と、それを解決する語義曖昧性解消について、分散表現についてである。次に、本論文で提案する語義の分散表現を利用した語義曖昧性解消の手法を説明する。最後にその手法を用いた実験と、それを踏まえた考察を述べる。

第2章　多義語とその曖昧性

2.1　多義語

我々が普段使っている「日本語」や「英語」などの、人間どうしの意思疎通に使われる言葉を自然言語という。自然言語で使われる単語には、複数の意味を持つ単語がある。そのような単語を多義語といい、単語が持つ意味を語義という。

単語「意味」を例にとって、多義語について具体的に説明する。岩波辞書において「意味」を引くと、以下の記述がある。

> **2843-0-0-1** その言葉の表す内容。意義。「辞書を引けば―がわかる」
>
> **2843-0-0-2** 表現や行為の意図・動機。「どういう―でそんなことをしたのか」
>
> **2843-0-0-3** 表現や行為のもつ価値。意義。「そんな事をしても―がない」

このように、それぞれの語義に対して、その定義文と例文が記述してある。「その言葉の表す内容。意義。」が、一つ目の語義の定義文であり、「『辞書を引けば―が分かる』」の部分が例文である。

これを踏まえると、単語「意味」は三つの語義を持っていることが分かる。今後は便宜的に、単語の語義をそれぞれ語義1、語義2 、語義3のように呼ぶことにする。

2.2　語義の曖昧性

例えば、会話の中で、

「そんなことに意味は無い。」

という文が出てきたとする。日本語が話せる人であれば、この曖昧な「意味」がどの語義としての「意味」なのかが予想できるだろう。これは、我々が普段会話

をしたり文を読んだりすることによって、文脈から語義を分類する能力を感覚的に身につけているからである。

機械に感覚はないので、同じ事を行う場合には語義曖昧性解消 (Word Sense Disambiguation, WSD) を組み込む必要がある。

語義曖昧性解消は、多義語が文中に出現した際、それがどの語義を表しているのかを判別するタスクである。語義曖昧性解消を組み込むことで、機械も語義を判別することが可能になる。

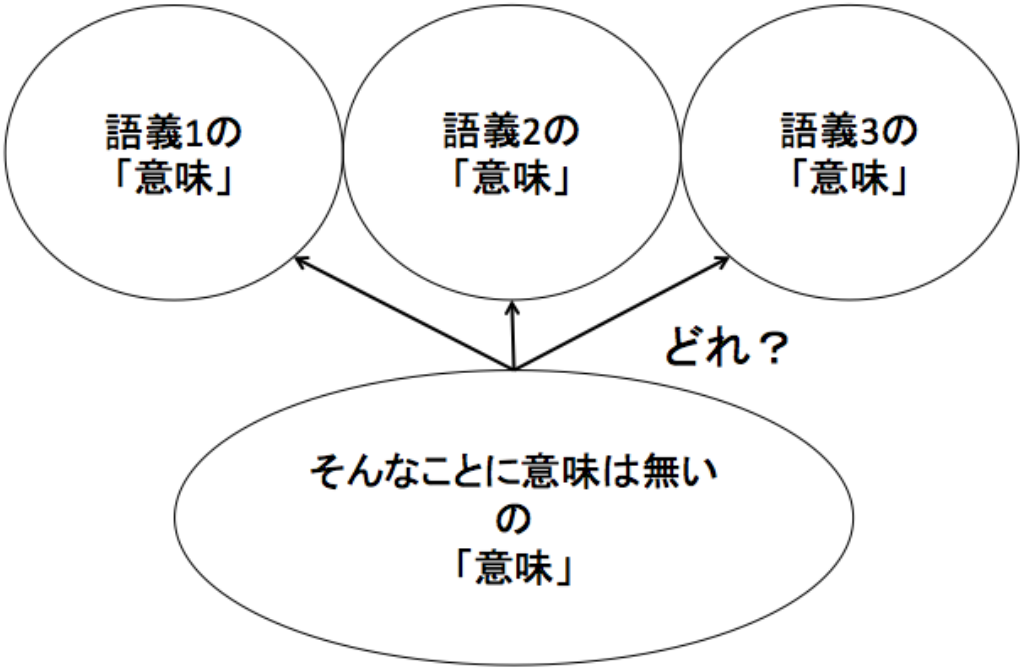

図 2.1: 語義の曖昧性

2.3 語義曖昧性解消

語義曖昧性解消を行うためには、分類器を定める必要がある。分類器は、多義語がどの語義であるのかを分類する人の感覚のようなものである。

分類器を定めるためには学習が必要である。学習には、教師あり学習手法や半教師あり学習手法、教師なし学習手法といった手法が利用される。

教師あり学習手法では、答えがあらかじめ書いてあるラベル付きデータというものを用い学習する。反対に教師なし学習手法では、答えが分からないラベル無し

データを用いる。半教師あり学習手法では、ラベル付きデータとラベル無しデータ共に扱う。

一般的に、語義曖昧性解消のタスクに対しては、教師あり学習手法が用いられる。

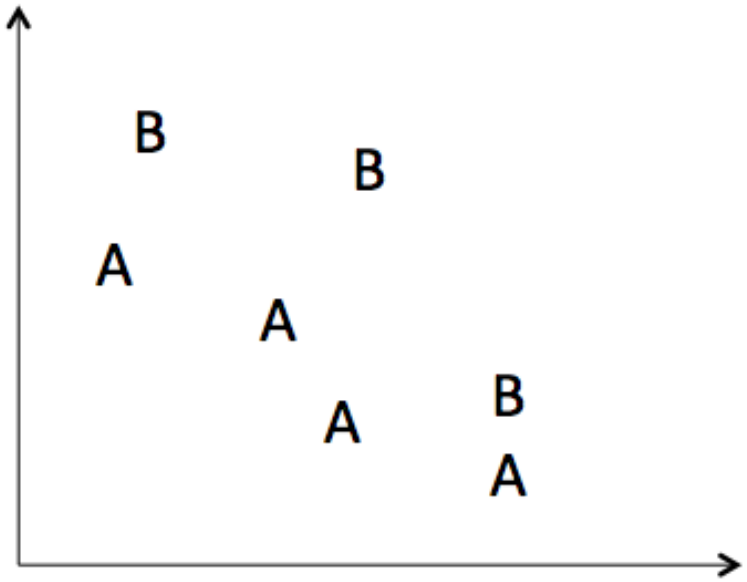

図 2.2: ラベル付きデータ

例えば図 2.2 のようなラベル付きデータを使って学習を行うとする。この例ではデータ A が語義 1 のデータ、データ B が語義 2 のデータである。

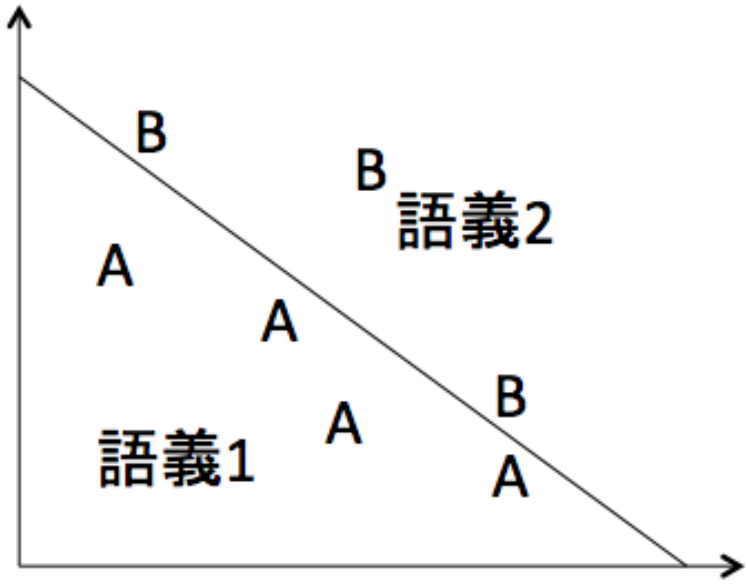

図 2.3: 教師あり学習後

図 2.3 は、そのようなデータに対して教室あり学習を行い、分類器を定めた状態である。棒線が識別の境界であり、この線を境に、語義 1 の側のデータを語義 1 へ、語義 2 の側に当てはまる単語を語義 2 へ分類する。

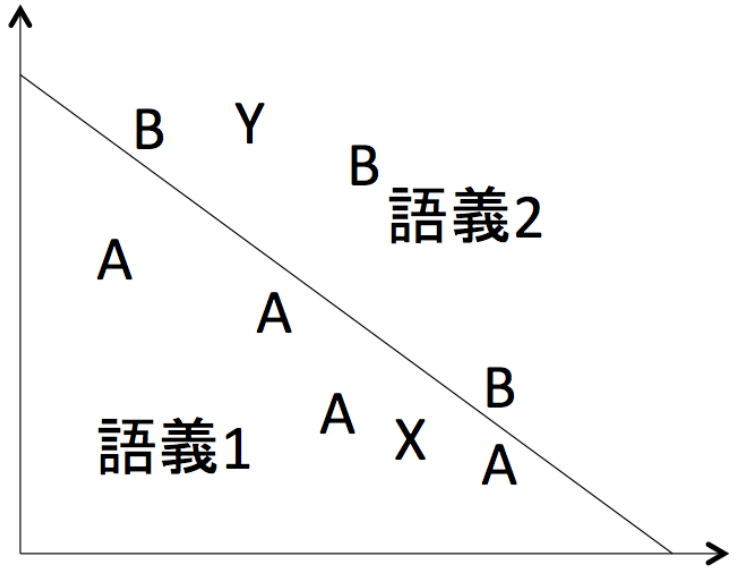

図 2.4: 未知のデータの分類

定めた分類器に対して図 2.4 のように未知のデータ X,Y を与えると、この分類器は X を語義 1、Y を語義 2 として分類する。

その他の語義曖昧性解消の手法として、lesk アルゴリズムがある。このアルゴリズムは辞書の定義文を使った知識ベースの古典的な手法である。定義文に出現する単語と、対象単語の周辺語が重複した数によって語義を識別する。

第3章　単語の分散表現

3.1　単語の数値化

人は知らない単語を見た時に、「どのような意味なのか」、「どのように使えばよいのか」、「どのような性質を持っているか」ということを、その単語が現れた文脈から予想したり、辞書を引いて調べることで学習することができる。

機械に単語を学習させる場合、単語を数値化する必要がある。従来は bag of words による高次元で疎なベクトルが利用されている。しかし、単語の距離がより正確に求められる事から、深層学習を利用した低次元で密なベクトルである分散表現を用いた手法がより良い表現であると考えられる。実際様々な自然言語処理のタスクに利用され、有効な結果を残している。

3.2　分散表現の構築

分散表現の構築には、[1] を用いる。word2vec は、skip-gram モデルを利用したテキストを処理するニューラルネットである。word2vec に文書のコーパスを与え、次元数などのパラメータを設定することで、その文書に出現する単語の分散表現を構築できる。

構築された単語の分散表現はベクトルであるため、足し引きできるという特徴がある。単語「王様」と「女王」、「男性」と「女性」を例に取り、その位置関係が図 3.1 のようになっているとする。

[1] https://code.google.com/p/word2vec/

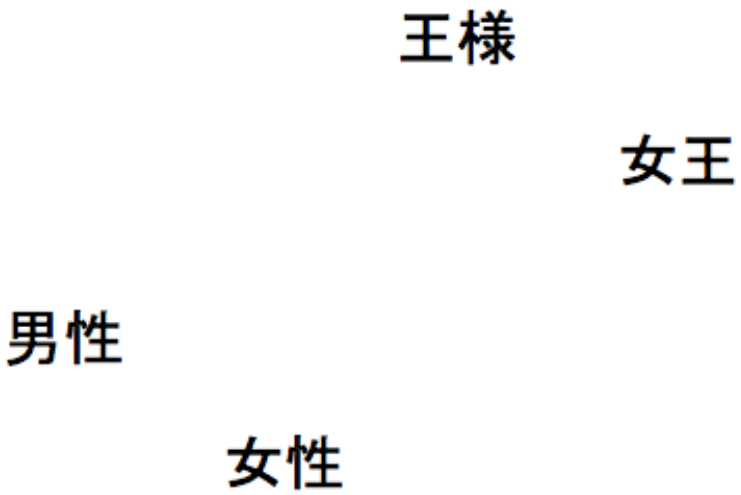

図 3.1: 王様と女王

これらは全てその単語の分散表現でありベクトルとなっている。この図からそれぞれの関係が、

$$\text{王様} - \text{男性} = \text{女王} - \text{女性}$$

のようになっていることが分かる。

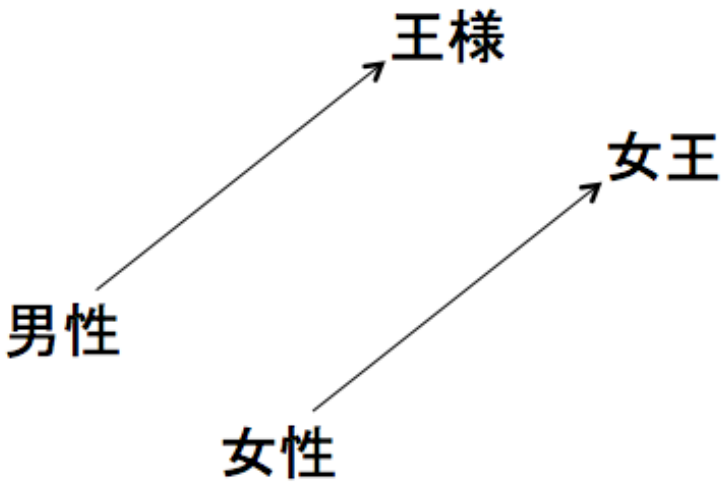

図 3.2: 王様-男性=女王-女性

「男性」から「王様」へ向かうベクトルと、「女性」から「女王」に向かうベクトルが等しくなっている。こう書くと少し分かりにくいが、要するに「王様」と「男性」という関係と、「女王」と「女性」という関係が等しいということである。また、この式を変形することで、

$$王様 - 女王 = 男性 - 女性$$

という形にもなる。

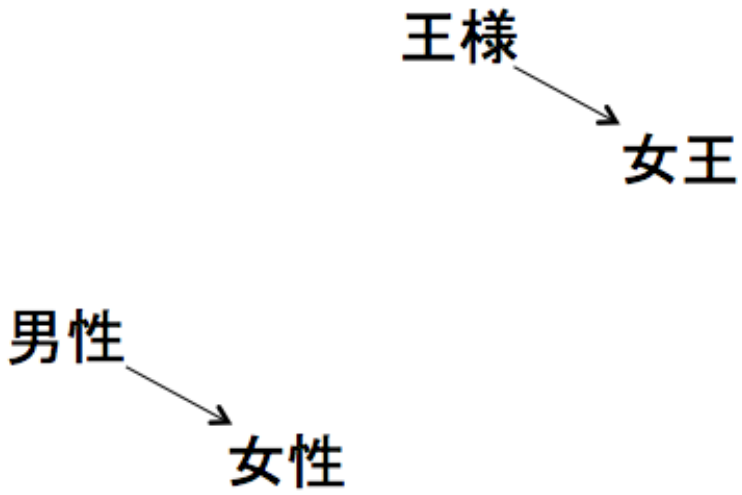

図 3.3: 王様-女王=男性-女性

これも同様に「王様」と「女王」という関係と、「女王」と「女性」という関係が等しいということである。さらにこの式を変形をすると、

$$王様 - 男性 + 女性 = 女王$$

という形になる。

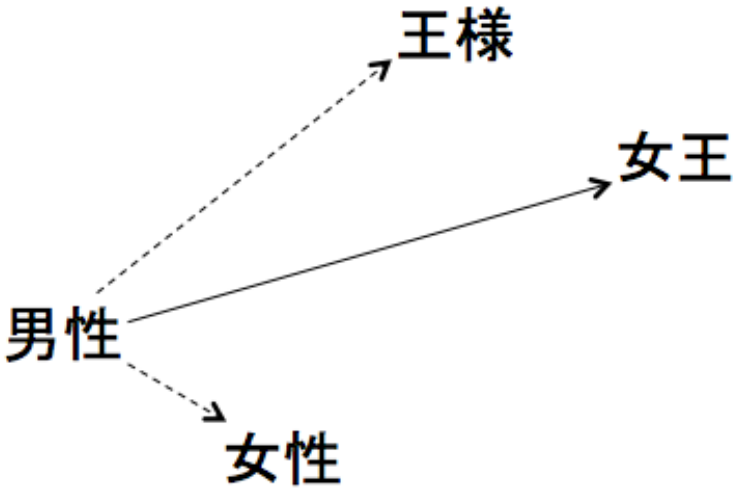

図 3.4: 王様-男性+女性=女王

「王様」から「男性」という要素を除いたものに「女性」を足すことで「女王」となるということである。

これらの関係は、人間である我々から見ても自然に理解できる。分散表現を利用すれば、このような単語どうしの演算が可能となる。

もう一つの例として、3 つの語義を持つ単語があるとする。その単語自信の分散表現を v、その単語の語義 1 の分散表現を v1、語義 2 の分散表現を v2、語義 3 の分散表現を v3 とすると、その関係は図 3.2 のようになると考えられる。

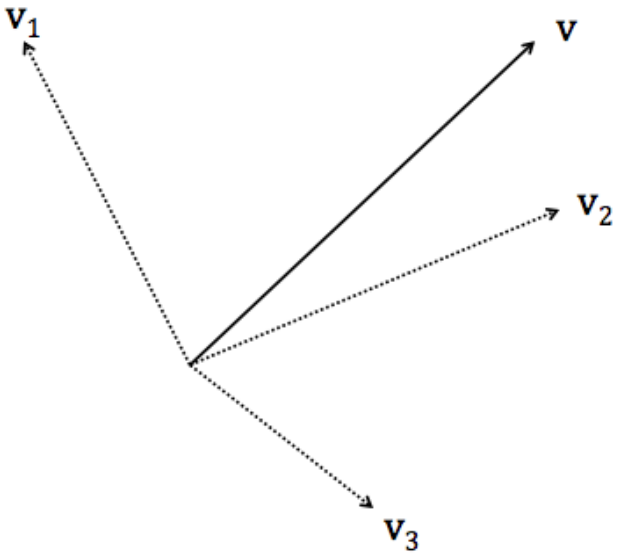

図 3.5: ある単語の分散表現

図 3.1〜図 3.5 では便宜的に 2 次元で表しているが、実際には 100 次元や 200 次元などに設定することが多い。

第4章　語義の分散表現による語義曖昧性解消

4.1　概要

一般的に、語義曖昧性解消のタスクに対しては教師あり学習手法が用いられる。しかし教師あり学習手法の場合、ラベル付きデータを用意しなければならないので、訓練データの作成コストが高い。そのため対象とする単語が限定されてしまい、あまり実用的ではないという問題がある。

本論文で提案する手法は、類義語の分散表現を語義の分散表現として利用する手法となる。分散表現の比較によって語義曖昧性解消が可能なため、語義の分散表現を求めることさえできれば、対象とする単語は限定されない。

4.2 提案手法

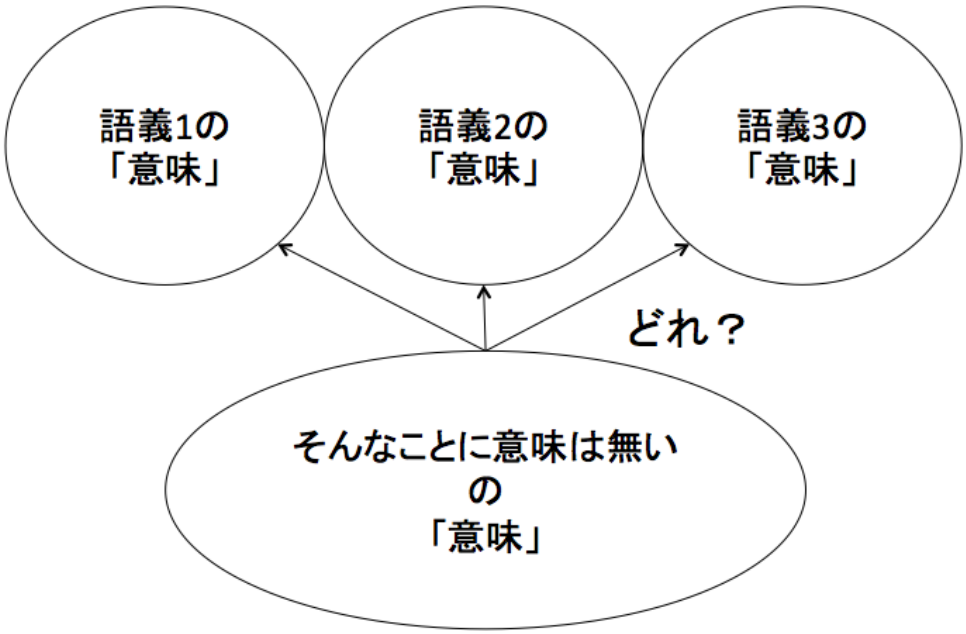

図 4.1: 本手法イメージ

対象単語の分散表現を $\boldsymbol{v}$ とし、その単語の各語義 $s_i(i = 1 \sim K)$ の分散表現を $\boldsymbol{v}_i$ とするとき、

$$\boldsymbol{v} = \sum_{i=1}^{K} \boldsymbol{v}_i$$

が成立しているとする。この関係式と s_i の語義を持つ類義語 w_i の分散表現 $\boldsymbol{u}_i$ を利用して、$\boldsymbol{v}$ から $\boldsymbol{v}_i$ を求めていく。

$\boldsymbol{u}_i \approx \boldsymbol{v}_i$ と考えられるため $\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$ とする。つまり、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

となる。最後に最小二乗法によりこの α_i を求めることで、$\boldsymbol{v}_i$ に $\boldsymbol{u}_i$ を近づけることができ、それによって単純に類義語を使うよりも、語義曖昧性解消の精度が上がると考えられる。

4.3　語義と類義語

ここからは、先程も例に上げた「意味」を使って、具体的に語義の分散表現を構築していく過程を示す。

単語「意味」は、岩波辞書において以下の 3 つの語義を持つ。

2843-0-0-1 その言葉の表す内容。意義。「辞書を引けば―がわかる」

2843-0-0-2 表現や行為の意図・動機。「どういう―でそんなことをしたのか」

2843-0-0-3 表現や行為のもつ価値。意義。「そんな事をしても―がない」

この定義文と例文を参考にしながら、類義語として利用する単語を集めていく。語義 1 の類義語として、「趣旨」と「内容」、語義 2 の同義語として、「目的」と「意図」と「動機」、語義 3 の同義語として「価値」と「重要性」という単語を集めた。

集めた単語が類義語になっているかどうかは、例文に当てはめてみなければわからない。そこで、ここからは岩波辞書の「意味」の例文に当てはめて、集めた単語が本当に類義語となっているのかを検証する。

語義 1 の例文に、「意味」と類義語として見つけた「趣旨」と「内容」を当てはめると、

「辞書を引けば意味が分かる」

「辞書を引けば趣旨が分かる」

「辞書を引けば内容がわかる」

となり、全て同じように通じる。

語義 2 の例文に、「意味」と類義語として見つけた「目的」と「意図」と「動機」を当てはめると、

「どういう意味でそんなことをしたのか」

「どういう目的でそんなことをしたのか」

「どういう意図でそんなことをしたのか」

「どういう動機でそんなことをしたのか」

となり、全て同じように通じる。
　語義 3 の例文に、「意味」と類義語として見つけた「価値」と「重要性」を当てはめると、

「そんな事をしても意味がない」

「そんな事をしても価値がない」

「そんな事をしても重要性がない」

となり、こちらも全て同じように通じる。

　以上のことから、集めた単語が類義語であることが分かった。よって、各語義に対する類義語を以下のように定める。

2843-0-0-1　「趣旨」「内容」

2843-0-0-2　「目的」「意図」「動機」

2843-0-0-3　「価値」「重要性」

4.4　類義語の分散表現

　語義曖昧性解消を行うために、語義毎の分散表現を求める必要がある。そのために、先程集めた類義語である、

2843-0-0-1　「趣旨」「内容」

2843-0-0-2　「目的」「意図」「動機」

2843-0-0-3　「価値」「重要性」

を利用する。語義の分散表現 $\boldsymbol{v}_i$ の代わりにこれらの類義語の分散表現 $\boldsymbol{u}_i$ と、その重み α を利用し、

$$\boldsymbol{v}_i = \alpha_i \boldsymbol{u}_i$$

として使用する。この $\boldsymbol{u}_i$は、集めた同義語から選出した単語の平均となる。例えば、語義 2 の同義語として「目的」と「意図」を使用するとする。それぞれの分散表現を $\boldsymbol{s}_1$、$\boldsymbol{s}_2$とすると、

$$\boldsymbol{u}_2 = \frac{\boldsymbol{s}_1 + \boldsymbol{s}_2}{2}$$

のようになるということである。

単語の分散表現は、BCCWJ コーパスから word2vec を用いて構築しておく。次元数は 100 とする。

4.5 語義の分散表現の重み

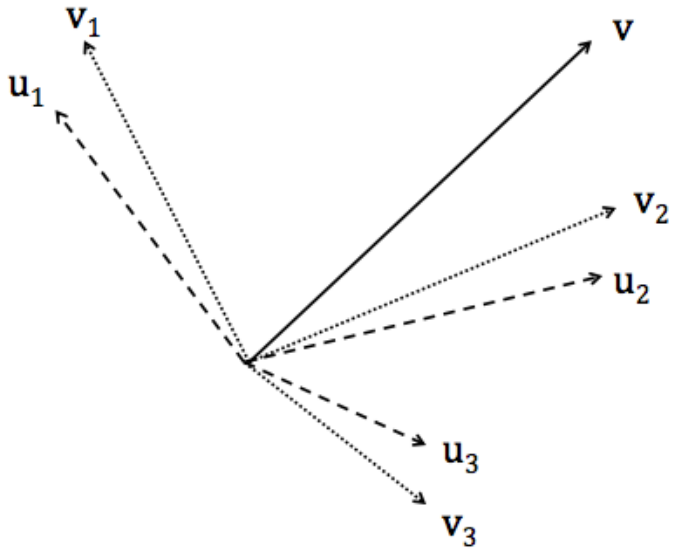

図 4.2: $\boldsymbol{v} \approx \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3$

図 4.2 は、単語「意味」の分散表現 $\boldsymbol{v}$ とその語義の分散表現 $\boldsymbol{v}_i$、類義語の分散表現 $\boldsymbol{u}_i$ の関係を、便宜的に 2 次元のベクトルで表したものである。$\boldsymbol{v}$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ 及び $\boldsymbol{u}_3$ が既知のベクトルであり、これらから $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ 及び $\boldsymbol{v}_3$ を構築する。

「序論」で述べたように、

$$\boldsymbol{v} = \sum_{i=1}^{K} \alpha_i \boldsymbol{u}_i$$

といえる。実際にはベクトルの大きさを揃えるため、正規化を行う必要がある。ベクトル $\boldsymbol{u}_n$ の要素を $\boldsymbol{u}_{ni}$ のように表すと、次元数は 100 なので、式は

$$u'_{ni} = \frac{u_{ni}}{\sqrt{\sum_{i=1}^{100} u_{ni}}}$$

となる。さらに、「意味」の語義は三つであるため $K = 3$ となる。それを踏まえると

$$\boldsymbol{v} = \sum_{i=1}^{3} \alpha_i \boldsymbol{u'}_i$$

のようになる。この α は、実際に単語が使用された際のそれぞれの語義の重みとなる。

この α を求めるために、最小二乗法を用いる。まず、

$$\alpha_1 \boldsymbol{u'}_1 + \alpha_2 \boldsymbol{u'}_2 + \alpha_3 \boldsymbol{u'}_3 - \boldsymbol{u'} = 0$$

といえる。この式を変形すると、

$$\alpha_1^2 |\boldsymbol{u'}_1^2| + \alpha_2^2 |\boldsymbol{u'}_2^2| + \alpha_3^2 |\boldsymbol{u'}_3^2| + 2\alpha_1\alpha_2 |\boldsymbol{u'}_1 \boldsymbol{u'}_2| + 2\alpha_2\alpha_3 |\boldsymbol{u'}_2 \boldsymbol{u'}_3| + 2\alpha_3\alpha_1 |\boldsymbol{u'}_3 \boldsymbol{u'}_1|$$
$$-\alpha_1 |\boldsymbol{u'}_1 \boldsymbol{u'}| - \alpha_2 |\boldsymbol{u'}_2 \boldsymbol{u'}| - \alpha_3 |\boldsymbol{u'}_3 \boldsymbol{u'}| + \boldsymbol{u'}^2 = 0$$

となる。上の式を $f(\alpha_1, \alpha_2, \alpha_3)$ とし、これに対し $\alpha_1, \alpha_2, \alpha_3$ に関する偏微分を行うと、

$$\frac{\partial f}{\partial \alpha_1} = 2\alpha_1 |\boldsymbol{u'}_1{}^2| + 2\alpha_2 |\boldsymbol{u'}_1 \boldsymbol{u'}_2| + 2\alpha_3 |\boldsymbol{u'}_3 \boldsymbol{u'}_1| - |\boldsymbol{u'}_1 \boldsymbol{u'}| = 0$$

$$\frac{\partial f}{\partial \alpha_2} = 2\alpha_2 |\boldsymbol{u'}_2{}^2| + 2\alpha_1 |\boldsymbol{u'}_1 \boldsymbol{u'}_2| + 2\alpha_3 |\boldsymbol{u'}_2 \boldsymbol{u'}_3| - |\boldsymbol{u'}_2 \boldsymbol{u'}| = 0$$

$$\frac{\partial f}{\partial \alpha_3} = 2\alpha_3 |\boldsymbol{u'}_3{}^2| + 2\alpha_2 |\boldsymbol{u'}_2 \boldsymbol{u'}_3| + 2\alpha_1 |\boldsymbol{u'}_3 \boldsymbol{u'}_1| - |\boldsymbol{u'}_3 \boldsymbol{u'}| = 0$$

となる。これらをまとめると、

$$\begin{pmatrix} |\boldsymbol{u'}_1{}^2| & |\boldsymbol{u'}_1 \boldsymbol{u'}_2| & |\boldsymbol{u'}_1 \boldsymbol{u'}_3| \\ |\boldsymbol{u'}_1 \boldsymbol{u'}_2| & |\boldsymbol{u'}_2{}^2| & |\boldsymbol{u'}_2 \boldsymbol{u'}_3| \\ |\boldsymbol{u'}_1 \boldsymbol{u'}_3| & |\boldsymbol{u'}_2 \boldsymbol{u'}_3| & |\boldsymbol{u'}_3{}^2| \end{pmatrix} \begin{pmatrix} 2\alpha_1 \\ 2\alpha_2 \\ 2\alpha_3 \end{pmatrix} = \begin{pmatrix} |\boldsymbol{u'}_1 \boldsymbol{u'}| \\ |\boldsymbol{u'}_2 \boldsymbol{u'}| \\ |\boldsymbol{u'}_3 \boldsymbol{u'}| \end{pmatrix}$$

という式が得られる。これを解くことで、$\alpha_1, \alpha_2, \alpha_3$ を求めることができる。

4.6　文脈の分散表現

実際に「意味」という単語が含まれる文脈の分散表現を求める。

文脈の分散表現を得るために、周辺語の分散表現を利用する。そのため、まず、対象単語「意味」の周辺の自立語を取り出す。

以下の例で考える。

「そんなことに意味は無い。」

上記の文からは、「そんな」、「事」、「無い」の三つの自立語を取り出すことができる。文脈の分散表現 $\boldsymbol{x}$ は周辺語の分散表現の平均に設定するので、これらの単語の分散表現をそれぞれ $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ とすると、$\boldsymbol{x}$ は次の式で得られる。

$$\boldsymbol{x} = \frac{\boldsymbol{x_1} + \boldsymbol{x_2} + \boldsymbol{x_3}}{3}$$

以上でこの文脈においての「意味」の分散表現を求めることが出来た。この分散表現と語義の分散表現を比較することで、類似度を割り出す事ができる。

4.7　語義曖昧性解消

ここまでに求めた値を用いて、語義曖昧性解消を行う。まず、類義語の分散表現と文脈の分散表現の類似度を求める。

ここではコサイン類似度を用いるため、分散表現 $\boldsymbol{x}$ を正規化しておく。次元数は 100 なので、式は、

$$x'_i = \frac{x'_i}{\sqrt{\sum_{i=1}^{100} x'_i}}$$

となる。新しく出来た文脈の分散表現と、類義語の分散表現をそれぞれ $\boldsymbol{x}'$ とすると、コサイン類似度は、

$$cos(\boldsymbol{x}', \boldsymbol{u}'_n) = \sum_{i=1}^{100} x'_i u'_{ni}$$

となり、さらに

$$c_n = cos(\boldsymbol{x}', \boldsymbol{u}'_n)\alpha_n$$

として重み付けをする。この c_n が最も大きい n が識別結果となる。

第5章　実験

5.1　概要

この実験の目的は本論文で述べた手法の正解率を割り出すことである。また、最小二乗法によって求めた重み付けの有効性についても調べる。

実験は単語「意味」を使った語義曖昧性解消を行う。テストデータとして、SemEval-2 の日本語辞書タスクの、「意味」のテストデータを用いた。[1] 重み付けの有効性を調べるために、重み付けをした場合としていない場合共に行う。

また、同義語の分散表現を利用する場合、同じ類義語でも、選択した単語やその組み合わせによって結果が大きく変わることがある。そのためただ全ての類義語を用いるのではなく、全ての組み合わせを試す。そして最も結果の良かった組み合わせと、最も悪かった組み合わせ、それぞれの正解率を割り出す。

5.2　結果

テストデータ 50 のうち、文脈の分散表現が構築できたのは 39 であった。そのデータと便宜上の番号を表 5.1 に記す。

[1] bag of words を使った手法で単語「意味」の語義曖昧性解消を行った際の正解率は、約 38 %である。

番号	周辺語
1	減 どう いう
2	良い 悪い 意味
3	言う 改革 ほとんど ない いう こと
4	目 以外 ある
5	どう いう
6	DH 呼び名 教え ください
7	念書 わから
8	脂肪 落とす いう 二の腕 細く
9	対策 し あまり ない
10	三十 また 恋愛
11	全く 違う 思い
12	どう いう ある 知り
13	大 天下 し 天下 元
14	プレ どう いう
15	先 言う 分かり 言う
16	承認 し OK
17	する 北米 インディアン 言葉
18	しゃらくせ〜 どう いう
19	害する もの 言う カラス
20	誠 まこと 語源 教え ください
21	相手 尊重 する 自己 防衛
22	自己 防衛
23	根 持っ なく 保管 し
24	高い こと する
25	開き 科学 技術
26	そう いっ
27	問う こと ある
28	問う こと 見出せる はず ない
29	こと 教え ない
30	やら あまり ない
31	変え ない マザー
32	使い 新しい いう ある
33	なっ いう
34	もどる そう いう ヌーヴェル
35	不明 問い合わせ 担当
36	さ ない いう 割り切り
37	五十 狭い セックス とどまら
38	実 ある
39	単語 通じる

表 5.1: 実験に使用するデータ

重み付けをして語義曖昧性解消を行った結果は、

2843-0-0-1 「趣旨」

2843-0-0-2 「目的」「意図」「動機」

2843-0-0-3 「重要性」

の組み合わせで最高の 59.0% 、

2843-0-0-1 「内容」

2843-0-0-2 「目的」「意図」

2843-0-0-3 「価値」「重要性」

の組み合わせで最低の 12.8% となった。

重み付けをせず語義曖昧性解消を行った結果は、

2843-0-0-1 「趣旨」

2843-0-0-2 「目的」「意図」「動機」

2843-0-0-3 「重要性」

の組み合わせで最高の 48.7% 、

2843-0-0-1 「趣旨」

2843-0-0-2 「目的」「含意」

2843-0-0-3 「価値」「重要性」

の組み合わせで最低の 12.8% となった。

この結果に対応する番号は、表 5.2 である。

番号	重み付け			
	有り		無し	
	最高	最低	最高	最低
1	正	誤	正	誤
2	誤	誤	誤	誤
3	誤	誤	誤	誤
4	誤	誤	誤	誤
5	正	誤	誤	誤
6	正	誤	正	誤
7	正	誤	正	誤
8	正	誤	誤	誤
9	誤	誤	誤	誤
10	正	誤	正	誤
11	正	誤	正	誤
12	正	誤	誤	誤
13	正	正	正	誤
14	正	誤	正	誤
15	正	誤	正	誤
16	正	誤	正	誤
17	誤	誤	誤	誤
18	正	誤	誤	誤
19	誤	誤	誤	誤
20	正	誤	正	誤
21	誤	正	誤	正
22	誤	正	誤	誤
23	正	誤	正	正
24	正	正	正	正
25	誤	誤	誤	誤
26	正	誤	正	誤
27	正	誤	正	誤
28	正	誤	正	誤
29	誤	誤	誤	誤
30	誤	誤	誤	誤
31	誤	誤	誤	誤
32	誤	誤	誤	誤
33	誤	誤	誤	誤
34	正	誤	正	誤
35	正	誤	正	誤
36	誤	誤	誤	正
37	誤	誤	誤	誤
38	正	誤	正	誤
39	正	正	正	正

表 5.2: 実験結果

第6章　考察

ここで構築した語義の分散表現が適切であるかどうかの評価は難しい。ただし実験では重み付けが有効であった。重み付けを行わないというのは単に類義語との類似性から語義を判定していることに対応する。また重み付けを行うというのは語義の分散表現を求めていることに対応する。このことから考えると、得られた語義の分散表現が類義語の分散表現以上には適切であったと考えられる。

本手法の問題点としては類義語を見つけることの困難性がある。ここで題材とした単語「意味」は各語義に対して類義語が見つけられたが、このような単語は稀である。また、今回集めた類義語においても、似ているだけで完全に一致しているとは言い難い。「内容」にしても、「目的」にしても多義語なので、「意味」と共通する語義を持ってはいるが、全く同じであるとは言えない。

語義の分散表現を見つける場合、語義の類義語が見つかれば、それは大きな手がかりとなるが、それは困難である。そのため、辞書の例文から文脈の分散表現を構築していく方向が現実的な手法と考えている。今後は辞書の例文を利用する方法を検討したい。

第7章　結論

本論文では、類義語を利用した語義の分散表現の構築方法を提案した。そして構築した分散表現を利用して、教師なしの語義曖昧性解消を行った。

実験では SemEval-2 の日本語辞書タスクでの単語「意味」のテストデータを用いて、重み付けをした場合（語義の分散表現を利用）としていない場合（類義語の分散表現を利用）との正解率を求めた。それぞれの最高値は、重み付けをした場合 59.0%、しなかった場合 48.7%となり、語義の分散表現が適切に構築されたと考えられる。

本手法の問題点は語義の類義語を見つけることが困難なことである。今後は辞書の例文を利用する方法を検討したい。

謝辞

本研究を進めるにあたって、様々な場面でお力添え頂いた新納浩幸教授に感謝を致します。

参考文献

[1] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A Unified Model for Word Sense Representation and Disambiguation. In *EMNLP-2014*, pp. 1025–1035, 2014.

[2] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58, 2007.

[3] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *EMNLP-2014*, pp. 1059–1069, 2014.

[4] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. On SemEval-2010 Japanese WSD Task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.