動詞の意味の曖昧性についての考察

宮田, 光樹
九州大学工学部電気情報工学科

鈴木, 孝彦
九州大学情報基盤研究開発センター

廣川, 佐千男
九州大学情報基盤研究開発センター

出版情報：情報処理学会研究報告. 2013 (A-4-2), pp.1-8, 2013-03-14. 情報処理学会九州支部
バージョン：

情報処理学会研究報告
IPSJ SIG Technical Report

動詞の意味の曖昧性についての考察

宮田光樹†1　鈴木孝彦†2　廣川佐千男†2
†1 九州大学工学部電気情報工学科　†2 九州大学情報基盤研究開発センター

概念辞書 WordNet の動詞には、複数の意味(曖昧性)を持つものが多い。この曖昧性を定量化し、語の難しさとどのような関連があるかを調べた。語の難しさの指標として、日本語能力試験(JLPT)の出題想定レベルを用いて分析した結果、難易度の高い動詞の方が難易度の低い動詞より曖昧性が低いことが分かった。また、文中の動詞の出現についての動詞の曖昧性解消に WordNet を利用する方法を提案し、単語の難易度と曖昧性解消性能の関係について考察した。

Consideration on Word Sense Ambiguity of Verbs

KOUKI MIYATA[†1]　TAKAHIKO SUZUKI[†2] SACHIO HIROKAWA [†2]

The concept dictionary WordNet contains many verbs with two or more meanings, i.e., ambiguity. This paper proposes a formulation of the ambiguity and analyzes if there is any relationship between the degree of ambiguity of a verb and the difficulty of the verb. As an index of the difficulty of a word, the level of Japanese Language Aptitude Test (JLPT) is used. The paper considers possibility of disambiguation of verbs.

1. はじめに

自然言語処理における重要な問題の一つに、多義語の意味的曖昧性の解消(Word Sense Disambiguation:WSD)がある[10]。多義語とは多くの意味を持っている語である。例えば「振り返る」が現れる次のような２つの分を考える。

文１：背後から声をかけられ「振り返る」と～
文２：この一年を「振り返る」

文１と文２では、「振り返る」という語が異なる意味で使われている。単語の意味を列挙したものを Sense Inventory(意味目録)という。

「振り返る」の Sense Inventory(意味目録)は次の２つから成り、文１、文２の意味はそれぞれ１番目、２番目に該当する。

振り返る１：後方へ顔を向ける。振り向く。
振り返る２：過ぎ去った事柄を思い出す。回顧する。

WSD は、文中に出現する多義語に、Sense Inventory のある一つの項目を対応付けることとみなすことができる。

WSD の手法は機械翻訳などのアプリケーションで非常に有用である。たとえば日本語の「振り返る」という動詞のもつそれぞれの語義は、ほかの言語では別々の動詞（英語では look back, review など）によって表されているためである。

著者は、この WSD の中でも動詞の多義性解消に着目した。動詞に注目したのにはいくつか理由がある。第一に、名詞よりも動詞の方が日常で使われている語数が少なく、その分多くの意味で使われることになり多義語の問題が強く表れると推定されるからである。第二に、WSD に関する論文では動詞を扱ったものは少なく、ほとんどは名詞を扱っているためである。このことから名詞よりも動詞の多義性解消の方が複雑であることも推定できる。第三の理由に、名詞と動詞の多義性を比較することで新しい知見が得られると期待されることがある。

また、近年大規模なオントロジー[4]の有用性が情報検索やデータ統合などの分野で注目されている。この大規模オントロジーは WSD にも有用であると考えられている。日本語の大規模オントロジーには、日本語 WordNet[13]、日本語語彙大系[14]などがある。

この大規模オントロジーの中で、本論文では WordNet について扱っている。WordNet では単語は Synset と呼ばれる概念（意味）毎に分類されている。また、２つの Synset の間には上位語、下位語などの繋がりもあるため類義語や上位語を探す際に用いることができる。今回は WordNet を使うことで多義語であるかどうかの判別を試みた。WordNet について詳細は第２章に記述する。WordNet のこのような特徴から、曖昧性の解消に有用であると考え、動詞多義語の意味判定に用いて有効性を確かめることにした。

また本研究では日本語の文章を対象とすることにした。それは、日本語を母語とする著者の感覚から、日本語の方が英語よりも曖昧な表現が多いと推察したからである。また、インフォーマルな記述が含まれることから、動詞の曖昧性がより顕著に現れるのではないかと推察し小説の文章を研究の対象に使用することにした。本論文では翻訳文章がある英語の小説の中でデータの収集が比較的簡単で内容が適切であるものとして、著者コナン・ドイルのシャーロック・ホームズシリーズを扱っている。英語の翻訳文章を使った理由に関しては第７章に記述する。

WSD は未解決の研究課題であり、本研究の範囲では実用的な規模の動詞の WSD システムを構築して評価するのは困難である。そこで、まず単語の難易度と多義性にはどのよ

うな関係があるかを調べることにした。単語の難易度は日本語能力試験(JLPT)を参考にして決定し、難易度別に単語のリストを作成した。

WordNet を使用し、そのリスト中の単語の持つ概念(Synset数)等を比較することによって、語彙の難易度によって多義語にどのような差が出てくるかを調べた。動詞と名詞について同様な調査を行い、本論文のテーマである動詞の多義性にどのような特徴があるかを比較した。

また、多義語の曖昧性を解消するために本論文では WordNet を使って多義語の各概念(Synset)で類義語の集合をとり、それを使って曖昧性の解消を目指した。各 Synset 別に作成した類義語の集合を用いて、文章中で集合に含まれる単語 w が出てきたらその周りの単語を収集し、それを周辺単語集合 Near(w)とした。そして、文章中に新たに出てきた多義語の周辺の単語と、この Near(w)を比べることによってその多義語の意味を判定するのが、本論文の最終目標である。

第２章では本研究で扱ったデータの収集に関する説明を行う。第３章では、単語の難易度による多義語の概念数の違いを比べる手法と結果について述べる。第４章では周辺単語集合 Near の構築法、Near を用いた多義性解消の手法、それに関する研究結果と評価を述べる。最後に第５章で本研究のまとめと今後の課題について述べる。

2. 関連研究

2.1 オントロジーに関する研究

オントロジーは、人工知能を始めとする情報系分野では「概念化の明示的仕様」と定義されている。言葉を階層構造等で体系化したものである。

大規模オントロジーには、日本語単語については、日本語語彙体系、EDR 電子辞書、日本語 WordNet、英単語については WordNet などが代表的である[5][8]。この中で本論文では大規模オントロジーの一つである日本語 WordNet を扱っている。[5]ではこれらのシソーラスの性能評価を行っている。

しかし、これらのオントロジーは手動で構築されているため、構築コストが大きく、膨大なコストがかかる、保守や更新が困難という問題がある。そのため近年では[8]のような（半）自動的にオントロジーを構築する手法、方法論、ツールなどの研究開発も進められている。[8]では、語彙網羅性、即時更新世に優れた、ユーザ参加型の Web 上の百科事典である Wikipedia を用いてオントロジーを構築を目指している。

2.2 WordNet を使った Classification

WordNet は Classification にもよく使用される。[5][7][11]では WordNet と機械学習を用いることによって Classification の精度を上げている。また、[7]では文章中の高頻出語と中頻出語の共通概念を抽出し、これを SVM(Support Vector Machine)の特徴語に使用することによってテキスト分類の精度を上げている。この共通概念を抽出する際に、WordNet の構造を活かして二つの単語に共通する上位語を見つけることで共通概念を見つけている。また、[11]は文章分類の学習データとしてとして WordNet から同義語を取り出して使用し、[5]は学習データとして上位語を取り出して使用して分類制度を向上させている。

2.3 動詞の多義性解消

[9]では、WordNet などの電子化辞書やシソーラス等を使用せず、コーパスだけで動詞の多義性解消を行っている。動詞が多義語かどうかの判定方法としては、本論文では WordNet を使用して分類しているのに対し、[9]ではコーパスのみを使って判定している。具体的には、動詞を動詞と共起して現れる名詞を軸にすることによりベクトル化し、その動詞ベクトルを名詞の軸に従い分割し、類義語とのクラスタの偏差を比較することによって決定している。

また曖昧性の解消に関しては、本論文では WordNet を使って各概念（Synset）別に Synset に含まれる類義語と共起して現れる語の集合を収集し、文中に現れる多義語と共起して現れる語を比較することにより曖昧性を解消している。それに対し、[9]では多義語とその多義語の意味を表す動詞と共起して現れる名詞の集合を用いて、文中に現れる多義語の曖昧性を、その語と共起する名詞と比較することにより解消している。

3. WordNetについて

WordNet[15]は、英語の大語彙データベースである。名詞、動詞、形容詞、副詞が「Synset」という異なる概念を表す単位によってグループ化されている。 Synset 同士は、概念的な意味と語彙の関係によって連結されている。そのため、類義語や上位下位語など意味的に関連している概念にブラウザ上で簡単にたどり着くことが出来る。WordNet は基本的に無償でデータの入手、利用が可能になっている。

WordNet は、意味に基づいて類義語を一つのグループにまとめる点で、表面的に通常のシソーラスと類似している。しかし、いくつかの重要な違いがある。 まず、WordNet は単語の文字列だけではなく、単語の特定の意味に関してリンクを張っている。 結果として、WordNet 上のネットワーク内で互いに近接して発見された単語は意味的に関係がある。 次に、類義語辞典で単語のグルーピングが類似性を意味する以外の明示的なパターンに従わないのに対し、WordNet は、単語間の意味的関係をラベル付けしている。

本研究では WordNet3.0 を基に作られた日本語 WordNet を使用している。

日本語 WordNet[13]は独立行政法人情報通信研究機構（NICT）が、大規模かつどなたでも入手できる日本語の意味辞書を開発することを目的とし作成したものである。

NICT がプリンストン大学で開発された Princeton WordNet（英語 WordNet）[15]や、Princeton WordNet とヨーロッパの EuroWordNet 協会[16]が推進する Global WordNet Grid[16]に着想を得て開発した。

日本語 WordNet では、英語 WordNet の Synset に対応して日本語が付与されています。 しかし、まだ Princeton WordNet には存在しない日本語 Synset を付与する必要があり、また、英語 WordNet の Synset の階層構造を日本語シソーラスとして 利用するためには、現状の Synset を修正する必要もある。日本語ワードネットに収録された Synset 数や単語数、語義数は次ようになっている。

57,238 概念 (Synset 数)
93,834 words 語
158058 語義 (Synset と単語のペア)

4. 分析対象データ

4.1 日本語能力試験（JLPT）データ

JLPT の単語については個人のホームページ[17]から収集した。JLPT の出題基準に関する文献[2]と照らし合わせて確認したが、一部不整合の部分があるがほぼ一致した。不整合の部分はホームページのリストを優先した。このデータから全単語のリストを作り、そのリストから旧 JLPT の試験 1 級の語彙、2-4 級の語彙に分けて作成した。また Near の収集と曖昧性解消の対象として小説シャーロック・ホームズを使用した。

日本語能力試験は、日本国内および海外において、日本語を母語としない人を対象として日本語の能力を測定し、認定することを目的として行う試験である。現在の日本語能力試験の受験級は N1-N5 の 5 段階に分かれている。認定の目安は N5 が「基本的な日本語をある程度理解することができる。」で、N1 が「幅広い場面で使われる日本語を理解することができる。」とされており、数字が小さくなるごとに難易度が上がるようになっている。しかし現在の試験で使われている受験級毎の単語のリストは非公開（公式には存在しないとされている）なため、本論文では 2009 年まで使われていた 1~4 級で分けられた語彙を収集して使用した。1~4 級の難易度レベルの目安は以下のようになっていて数字が小さくなるにつれて難易度が上がっている。

- １級-高度の文法・漢字（2,000 字程度）・語彙（10,000 語程度）を習得。社会生活をする上で必要な、総合的な日本語能力。900 時間程度学習したレベル。
- ２級-やや高度の文法・漢字（1,000 字程度）・語彙（6,000 語程度）を習得。一般的なことがらについて、会話ができ、読み書きできる能力。600 時間程度学習し、中級コース修了したレベル。
- ３級-基本的な文法・漢字（300 字程度）・語彙（1,500 語程度）を習得。日常生活に役立つ会話ができ、簡単な文章が読み書きできる能力。300 時間程度学習し、初級コース修了したレベル。
- ４級-初歩的な文法・漢字（100 字程度)・語彙（800 語程度)を習得。簡単な会話ができ、平易な文、又は短い文章が読み書きできる能力。150 時間程度学習し、初級コース前半を修了したレベル。

JLPT の試験は２０１０年に改訂され、旧試験の２級と３級の間のレベルとして N3 が設立されたことが一番大きな違いである。N3 の認定の目安は「日常的な場面で使われる日本語をある程度理解することができる」となっている。

4.2 シャーロック・ホームズシリーズ

シャーロック・ホームズシリーズは、小説家アーサー・コナン・ドイルの作品で、シャーロック・ホームズと、友人で書き手のジョン・H・ワトスンの織り成す冒険小説の要素を含む推理小説である。1887 年から 1927 年にかけて、60 編（長編 4、短編 56）が発表された。

本論文では、青空文庫[18]からシャーロック・ホームズシリーズの翻訳文章を収集した。青空文庫は無償で利用できるインターネット図書館である。著作権の消滅した作品と「自由に読んでもらってかまわない」とされたものが、テキストと XHTML（一部 HTML）形式で揃えられている。また、今回は使用していないが、今後の研究で用いるために英語のシャーロック・ホームズの文章も収集した。そちらは「The complete Sherlock Homes Download」[19]を利用した。ここでは、無償で英語のシャーロック・ホームズの文章が PDF、ePUB、HTML、ASCII 形式で提供されている。今回収集した文章は以下の表通りである。「赤毛連盟」、「舌のねじれた男」「三枚の学生」は Near の収集に使用した。「蒼炎石」は曖昧性解消の対象として使用した。Near とそれを用いた曖昧性解消に関しては第６章に記述する。

作品名	原題	翻訳者	データ量 (KB)	文字数	使用目的
赤毛連盟	The Red-Headed League	大久保ゆう	46	23,328	Near の収集
舌のねじれた男	THE MAN WITH THE TWISTED LIP	大久保ゆう	42	21,009	Near の収集

三枚の学生	THE ADVENTURE OF THE THREE STUDENTS	大久保ゆう	29	14,356	Near の収集
蒼炎石	THE ADVENTURE OF THE BLUE CARBUNCLE	大久保ゆう	36	17,897	曖昧性解消

表１、使用したシャーロック・ホームズシリーズについて前章で述べた通り、無作為に抽出した 500 英語添削例のうち、全文を書き直したものを除いた 399 例を誤りパターン分類のために使用した。

5. 単語の難易度による多義性の違い

5.1 単語の難易度による多義性の違い

この実験の目的は JLPT の級で分けられた語彙の難しさと WSD の難しさにどのような関係があるかについて、予備的な調査を行うことにある。また、動詞と名詞で WSD の難しさに違いが出てくるかの確認も行っている。

WSD の困難さの予測指標として、(1)単語の意味の数（WordNet 中で単語が属する Synset の数：概念数）(2)類義語の数、および、(3)多義語の異なる意味の重複度（単語が属する異なる Synset 間で、類義語がどれだけ重複しているか）を採用した。

JLPT の難易度別に、動詞と名詞について上記の指標を調べた。作成した JLPT1(JLPT1 級語彙のリスト)、JLPT2-4(2-4 級語彙のリスト)を WordNet 上で名詞として扱われているものと、動詞として扱われているものに分ける。それらを JLPT1_n,JLPT1_v,JLPT2-4_n,JLPT_v とする。また、全単語を名詞と動詞に分けて作成したリストを、JLPY1-4_n,JLPT1-4_v とする。以下が行った実験の手順となる。

(1) Synset 数

作成したリストの各単語について Synset 数を求め、Synset 数の平均、標準偏差を求める。WordNet に単語を入力し検索すると、単語の属する概念数の数だけ Synset が出力される。それを単語の持つ Synset 数とする。Synset 数をリストの全ての単語について求め、その結果の平均と標準偏差を求める。この実験は LPT1_n, JLPT1_v, JLPT2-4_n, JLPT_v, JLPY1-4_n, JLPT1-4_v のリストについて行う。

(2) 基本統計量

Synset 中の word 数の平均、標準偏差を求める。各 Synset 中にはその概念に属する単語が入っている。その単語の数を各 Synset 中の word 数とする。単語のリストの各単語で検索した際出力される Synset を使って、Synset 中の word 数を求める。Synset 中の word 数をリストすべての単語について求め、その結果の平均と標準偏差を求める。この実験は LPT1_n,JLPT1_v,JLPT2-4_n,JLPT_v,JLPY1-4_n,JLPT1-4_v について行う。

(3) 重複度

多義語に対する Synset 中の、語の重複度を求める。またその平均、標準偏差を求める。各単語の Synset 中の word は単語中の他の Synset 中の word と重複しているものがある。その重複度合について求める。具体的には、重複している単語数/ 重複無しの単語を含む全単語数の平均を求めることにより重複度を測る。ここで一つ例を挙げる。ある単語の Synset に含まれる単語の集合を Synsetn(v)とする。

Synset1(戻す)={返上, 還元, 返す, 返納, 戻す, 返却}
Synset2(戻す)={復旧,復する,還元,戻す,立て直す, 回復}
Synset3(戻す)={嘔吐, 吐く, 吐き出す, 戻す}

この場合 Synset1(戻す)の単語数は 6、Synset2（戻す）の単語数は 6、Synset3(戻す)の単語数は 4 である。また、Synset1（戻す）中で他の Synset と重複している単語は{還元、戻す}なので重複している単語数は２、また Synset2(戻す)の重複してる単語数は 2、Synset3(戻す)の重複している単語は１である。この結果から、重複度は 2+2+1/6+6+4=5/16 = 0.3125 となる。この重複度をリスト中のすべての単語について求め、その結果の平均と標準偏差を求める。この実験は LPT1_n, JLPT1_v, JLPT2-4_n, JLPT_v, JLPY1-4_n, JLPT1-4_v について行う。 このデータを使って T 検定、F 検定を行う。

5.2 結果と評価

単語を難易度別に分けてその Synset 数、単語数（類義語の数）、重複度（類義語の重複）の単語数（Synset 数）、合計、平均、分散、標準偏差を求めた。また、名詞と動詞についても別々に調べた。その結果を下の表に表示する。また、JLPT では動詞として扱われていなくても WordNet 内で動詞として扱われているものは動詞として扱った。

Synset 数

	単語数	合計	平均	分散	標準偏差
JLPT2-4（動詞）	1289	5860	4.546	13.55	3.681
JLPT1（動詞）	906	2953	3.259	6.157	2.481
JLPT1-4（動詞）	2195	8813	4.015	10.90	3.301
JLPT2-4（名詞）	2556	8717	3.410	7.029	2.651
JLPT1（名	1732	5248	3.030	5.443	2.333

詞)					
JLPT1-4（名詞）	4288	13965	3.257	6.423	2.534

表２、各単語について WordNet 中の Synset 数の比較

表２は各単語の Synset 数の合計、平均、分散、標準偏差を JLPT1、JLPT2-4、JLPT1-4(JLPT の全ての単語)毎に動詞と名詞について求め比較したものである。
難しい単語が多く含まれる JLPT1 の方が synset 数の平均は多くなると推測していたが、逆に JLPT2-4 の方が多い結果となった。

単語数

	Synset 数	合計	平均	分散	標準偏差
JLPT2-4（動詞）	5860	66881	11.41	128.1	11.32
JLPT1（動詞）	2953	34198	11.58	129.7	11.39
JLPT1-4（動詞）	8813	101079	11.47	128.6	11.34
JLPT2-4（名詞）	8717	81336	9.331	90.07	9.491
JLPT1（名詞）	5248	49495	9.431	86.34	9.292
JLPT1-4（名詞）	13965	130831	9.368	88.67	9.417

表３、各単語の各 Synset について Synset に属する類義語の数の比較

表３は各単語の各 Synset に属する類義語の数の合計、平均、分散、標準偏差を、表２と同様に単語の難易度別に動詞、名詞について求め、比較したものである。
JLPT1,JLPT2-4 の間では単語数の平均、標準偏差ともにほとんど差は見られない。動詞と名詞を比べるとやや動詞の方が平均が高いように思える。

重複

	単語数	合計	平均	分散	標準偏差
JLPT2-4（動詞）	1289	514.9	0.3995	0.07806	0.2794
JLPT1（動詞）	906	310.5	0.3427	0.08387	0.2896
JLPT1-4（動詞）	2195	825.4	0.3761	0.08123	0.2850
JLPT2-4（名詞）	2556	857.6	0.3355	0.07650	0.2766
JLPT1（名詞）	1732	547.4	0.3161	0.08104	0.2847
JLPT1-4（名詞）	4288	1405	0.3277	0.07842	0.2800

表４、各単語の各 Synset に属する類義語の重複度の比較

表 4 は各単語の各 Synset に属する類義語の重複度の合計、平均、分散、標準偏差を、表２，３と同様に単語の難易度別に動詞、名詞について求め、比較したものである。
重複度の平均は動詞・名詞ともに JLPT2-4 の方が大きいように見える。また、動詞の方が名詞よりも重複度の平均は大きいように見える。
この結果が統計的に差があると言えるかを確かめるために t 検定、F 検定を行った。Synset 数、単語数、重複度にの分布ついて、動詞と名詞それぞれ JLPT1、JLPT2-4 の二つの分布を使い、t 検定、F 検定を行なった。また、同様に JLPT の全ての単語について動詞、名詞の二つの分布についても検定を行った。結果は以下の通りである。
Synset 数

	F 値	分散が同じ確率	t 値	自由度	両側確率	両側有意
動詞	2.203	8.405E-36	9.780	2189	0.000%	○
名詞	2.203	4.453E-9	4.96	4007	0.000%	○
動詞名詞	2.203	1.755E-48	9.43	3552	0.000%	○

表５、Synset 数の分布についての t 値・F 値検定の比較

表５は Synset 数の分布について、JLPT1 と JLPT2-4 の二つの分布について、動詞と名詞それぞれで t 値・F 値検定を行い、また JLPT の全ての単語における動詞と名詞の二つ分布で t 値・F 値検定を行いそれを比較したものである。
各々の検定の流れと結果について下に表示する。
[動詞]分散が等しいという仮定が成り立たないので Welch の式を用いた。結果から２群の母平均に差があると言える。
[名詞]分散が等しいという仮定が成り立たないので Welch の式を用いた。結果から２群の母平均に差があると言える。
[動詞・名詞間]分散が等しいという仮定が成り立たないので Welch の式を用いた。結果から２群の母平均に差があると言える。
動詞・名詞共に Synset 数は有意に異なるという結果になった。このことから JLPT2-4 は JLPT1 よりも Synset 数は多いと言える。また動詞と名詞を比べた場合も Synset 数は有意に異なるという結果であった。このことからは名詞よりも動詞の方が Synset 数の平均は多いということがわかる。

・単語数（類義語の数）

	F 値	分散が同じ確率	t 値	自由度	両側確率	両側有意
動詞	2.203	0.6509	0.6500	8811	51.26%	×
名詞	2.203	0.04395	0.6500	11240	51.67%	×
動詞-名詞	2.203	1.428E-85	14.51	16210	0.000%	○

表６、単語数の分布ついての t 値・F 値検定の比較

表６は単語数の分布について、表５と同様に t 値・F 値検定を行いそれを比較したものである。各々の検定の流れと結果について下に表示する。

[動詞]分散が等しいという仮定が成り立つので通常の t 検定を行った。結果から２群の母平均に差があるとは言えない。

[名詞]分散が等しいという仮定が成り立たないので Welch の式を用いた。結果から２群の母平均に差があるとは言えない。

[動詞・名詞間]分散が等しいという仮定が成り立たないので Welch の式を用いた。結果から２群の母平均に差があると言える。

動詞・名詞共に、JLPT の難易度の違いによる有意な差は見られなかった。よって単語の難易度によって類義語の数に違いは無いとみなせる。しかし、動詞と名詞の間には有意な差がみられた。このことから動詞の方が名詞よりも Synset 中に含まれる類義語の数は多いとみなせる。

重複

	F 値	分散が同じ確率	t 値	自由度	両側確率	両側有意
動詞	2.203	0.8387	4.610	2193	0.000%	○
名詞	2.203	0.9057	2.230	4286	2.550%	○
動詞-名詞	2.203	0.1700	6.540	6481	0.000%	○

表７、重複度の分布についての t 値・F 値検定の比較

表７は単語数の分布について、表５、６と同様に t 値・F 値検定を行いそれを比較したものである。各々の検定の流れと結果について下に表示する。

[動詞]分散が等しいという仮定が成り立つので通常の t 検定を行った。結果から２群の母平均に差があると言える。

[名詞]分散が等しいという仮定が成り立つので通常の t 検定を行った。結果から２群の母平均に差があると言える。

[動詞・名詞間]分散が等しいという仮定が成り立つので通常の t 検定を行った。結果から２群の母平均に差があると言える。

動詞・名詞共に、重複度は有意に異なるという結果となった。このことから JLPT2-4 の方が重複度の平均は高いとみなせる。また動詞と名詞の間でも、重複度は有意に異なるという結果となった。このことから動詞の方が名詞よりも重複度の平均は高いとみなせる。

WordNet 上の単語の多義性の指標について、動詞／名詞間、JLPT1/JLPT2-4 間で、多くの場合統計的に有意な差があることがわかった。ここから、多義性と品詞および多義性と単語の難易度になんらかの関係があることを意味する。

特に、本研究の対象である動詞について、多義性の指標と単語の難易度の間に、大きな違いがあることがわかった。当初の予想とは異なり、難易度の高い動詞のほうが、低い動詞より、著しく Synset 数が少ない。これは、難易度の高い動詞に注目することで 日本語 WordNet を使った WSD に新しい知見が得られる可能性を示唆している。

6. 動詞の難易度との多義性解消性能

6.1 Near を用いた多義性解消

WSD では、目標とする多義性を持つ単語の近傍の単語の発生頻度を調べることが基本となる。「振り返る」の例では、近くに場所や方向を示す、「自宅」「後ろ」などの単語があれば「後方へ顔を向ける」の意味、時間や時代を表す「昨日」「昭和時代」などの単語があれば「回顧」の意味を表している場合が多い。このように近傍の単語は多義性解消の手がかりとなる。

目標とする単語の近傍にどのような単語が共起するか判断するためには、一般には、あらかじめ曖昧性を解消したコーパス等の例（学習データ）を多数準備し、そこから共起データベースを作成する[10]。しかし、本研究では、特別な学習例は用意せずに、WordNet の Synset 中に含まれる目標とする単語の類義語と一般のテキストを用いて共起データベースを作成した。例として「振り返る」では、第一の意味(Synset)に属する類義語「振り向く」、第二の意味の類義語「回想する」それぞれについて、一般の文章データ中の使用例を集め、それぞれの共起データベースを作る。

文章中に出現する、曖昧性を解消したい単語「振り返る」のそれぞれの例について、「振り向く」と「回想する」の共起データベースと近傍の語とを比較し、一致数が多いものをその「振り返る」の例の意味とする[9][12]。なお、本研究では「語の近傍」を「同一文内」と定義した。

WordNet を使って特定の単語 w から、w の各 Synset 別に類義語の集合を作成する。次に、文章 A 中で類義語の集合に含まれる単語が出現したら、その出現した単語を含む一文を１単語毎に区切りその単語を収集する（ただし出現した

単語そのものは収集しない)。それを w の各 Synset 別に行い収集した単語の集合が Near である。単語 w、および文集合 A、概念インデックス id について、Near(A,W,id)={X|w∈Syn(id),X∈now(A,v)}として定義する。ただし、Sent(A,v)は A 中で v を含む文の集合(要素の重複を許す)、now(A,v)は、{u|u∈Sent(a,v)}を表す。この時 now(A,v)は v を含まない。また Near を収集する際、文中に現れる「や、などの記号は含まない。Near を収集するための文書集合については 4.1.2 で説明している。

曖昧性解消を行いたい文章中の単語(動詞)を t とする。t を含む文中の単語(t 自身を含まず、重複を許す)の集合 U(t)と Near(A,id1,w)…Near(A,idn,w)を使って曖昧性解消を行う。この時、U(t)は Near と同様に「や、などの記号は含まない。曖昧性解消を行う文書に関しては 4.1.2 で説明している。

U(t) ∩ Near(A,id1,w)/U(t) ∪ Near(A,id1,w),…, U(t) ∩ Near(A,idn,w)/U(t)∪Near(A,idn,w)を比較し、最も数値が大きかったものの Synset(idi)を単語 t の概念(単語の意味)とする。

単語 t が JLPT1 に含まれる場合と JLPT2-4 に含まれる場合について、別々に実験を行い結果の違いを評価した。

6.2 結果と評価

本研究の手法で曖昧性が解消できたか検証するには人間が手作業でする必要があり、全ての結果について検証するには量が多すぎたため、今回は曖昧性が解消された例を一つ抜き出して表示する。

表８は文 A に出現した[取り組む]について、本研究の手法を用いて曖昧性解消を行った結果である。また、日本語 WordNet 上で[取り組む]の各 Synset の概念を説明した文章が表９となっている。本手法を用いると文 A 中に現れた[取り組む]の意味は Synset4 の「懸命であるか非常な努力をする」となる。文 A を見ると各 Synset の中で Synset4 が最も適切な意味である。よって、文 A 中の[取り組む]において曖昧性が解消された。

また、JLPT1 と JLPT2-4 の間で曖昧性解消を行った際の難しさを比べるために、χ二乗検定を用いた。

> 文 A「僕らの今[取り組む]べき問題は、どのような道を経て、物色された宝石箱という始点から、トテナム・コート通りの鵞鳥の餌袋という終点に至ったかだ」

SS1	SS2	SS3	SS4	SS5	SS6	SS7
0	0	0.5076	0.7111	0	0.5534	0

表８、[取り組む]の曖昧性解消

SS1	何かを行い、実行するために行う
SS2	対立する風潮または勢力に打ち勝とうとする戦い
SS3	戦争行為または競争などにおいて敵対する
SS4	懸命であるか非常な努力をする
SS5	深く考える、研究する、または討論する
SS6	難局として受け入れる
SS7	戦う、または近距離で混乱した状態で争う

表９、[取り組む]の各 Synset の概念

χ二乗検定を行う際、U(t)∩Near(A,id1…n,w)を実測度数とする。また(U(t)∩Near(A,id1,w)+ U(t)∩Near(A,id2,w)+…U(t)∩ Near(A,idn,w))/ (U(t) ∪ Near(A,id1,w)+ U(t) ∪ Near(A,id2,w)+…U(t)∪Near(A,idn,w))を平均割合 m とする。これを使って U(t)∩Near(A,id1…n,w)*m を期待度数としχ二乗検定を行った。JLPT1 と JLPT2-4 のそれぞれの結果を比較したものが表１０になる。

	χ二乗分布の片側確率が 0.5 以下の割合	0.05 以下の割合
JLPT1	0.3256	0.1628
JLPT2-4	0.0775	0.02584

表１０、JLPT1,JLPT2-4 におけるχ二乗分布の比較

この結果から JLPT1 の方が曖昧性解消は上手くいっていると考えられる。

また JLPT1、JLPT2-4 についてχ二乗分布の片側確率の分布に関して、平均と標準偏差を求めた。

	合計	平均	分散	標準偏差
JLPT1	453.1	0.9008	0.05256	0.2293
JLPT2-4	28.09	0.6531	0.1496	0.3868

表１１、JLPT1,JLPT2-4 についてχ二乗分布の片側確率の比較

この結果を用いて t 検定 F 検定を行った結果、F 値:0.351337、分散が同じ確率:0.99996 となった。分散が等しいという仮定が成り立たつので通常の t 検定を行った。その結果、t 値:6.36、自由度:544 両側確率:0.00% 両側有意:有意となった。よって２群の母平均に差があると言えるので、統計的にも JLPT2-4 の方が平均が高いと言える。

この結果より、動詞の難易度によって曖昧性解消に差が現れることが分かった。すなわち、難易度の高い JLPT1 に属する動詞の曖昧性解消が上手くいっている。この結果は第５章で述べた、WordNet 中の指標と一致している。

JLPT1 の方が曖昧性解消が上手くいった理由としては、JLPT2-4 に属する簡単な単語は日頃多く使われることによって多くの意味に分かれているためだと考えられる。

7. まとめと今後の課題

第５章の結果より、単語の難しさと WSD には強い相関があることが分かった。また、動詞と名詞に関しても、単語

の意味の数(Synset 数)、単語の類義語の数、類義語の重複度の平均において、動詞の方が高いことが分かった。このことから名詞よりも動詞の多義性解消の方が複雑であると推察でき、動詞で WSD を行うことは重要であることが分かる。特に、本研究で示した、難易度の高い動詞の WSD が難易度の低い動詞の WSD よりも良い結果が出るという傾向は、今後研究のために重要である。

本論文では提案した曖昧性解消の手法について正解率を求めることはできなかった。正解率を求めるために手動で確認する必要があり、膨大な量により今回は出来なかったので、正解率を求める方法に関しては今後と課題の一つとする。

また、曖昧性解消の精度に関してもあまりいい結果は出なかった。原因としては、Near のサンプルが少ないこと、日本語 WordNet の完成度等が考えられる。また Near を収集する際に、ターゲットとする単語が出現する一文を持ってきているが、これを単語の近傍の数単語に絞ることによって精度の改善が見込まれる。サンプルに関しては量を増やすだけではなく、内容に関してももう一度吟味することが必要であると考えられる。また、今回は Near を収集する際に類義語しか使用していなかったが、今後は上位語下位語を使うことによって精度の向上を図ることも考えている。

また、今後の課題の一つに日本語と英語の WSD にどの程度差があるかの確認もある。また日本語 WordNet と英語の WordNet の完成度に差があると考えたのは、日本語 WordNet は英語の WordNet を基に作成されているからである。そのため将来的に日本語と英語の WSD の比較と日本語 WordNet と英語の WordNet の比較をするために、本論文では英語の文書を日本語に翻訳したものを使用している。

また、5.2 の[取り組む]の例からも感じられるように、WordNet は語の意味を細かく分けすぎているように思える。日本語 WordNet 以外にも日本語の大規模オントロジーは存在しており、日本語 WordNet 以外の大規模オントロジーを使用することによって WSD の精度に差が生じると推測されるため、使用するオントロジーによる精度の検証も今後の課題の一つとする。

また、最近、コーパスから意味的に近い語群の情報や、共起関係の情報などを抽出する研究が盛んに行われている[9]。そのため、今後はコーパスのみを用いた WordNet 等のデータを用いない曖昧性解消についてもアプローチしていく必要があると考えられる。

WSD は人工知能の分野でも難しいとされる未解決研究の一つであり、本研究でその中でも動詞の多義性解消は複雑であると分かった。今後は今回見つかった課題に取り組みつつ WSD の精度を向上させていく予定だ。

参考文献

1) Manabu Okumura,Kiyoaki Shirai,Kanako Komiya,Hikaru Yokono,On SemEval-2010 Japanese WSD Task,Information and Media Technologies,Vol.6(2011),No.3,pp.730-744
2) 国際交流募金,日本語国際教育協会,日本語能力試験出題基準,凡人社,230pp
3) 栗原伸一,入門統計学：検定から多変量解析・実験計画法まで,オーム社,336pp
4) 兼岩憲,記述論理と Web オントロジー言語,オーム社,192pp
5) 川島貴弘,石川勉,言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価,人工知能学会論文誌,AI20(2005),pp326-336
6) Sam Scott,Stan Matwin,Text Classification Using WordNet Hypernyms, USE OF WORDNET IN NATURAL LANGUAGE PROCESSING SYSTEMS: PROCEEDINGS OF THE CONFERENCE,pp38-44,1998
7) 猪野陽子,松井藤五郎,大和田勇人, WordNet からの共通概念抽出によるテキスト分類,日本ソフトウェア科学会第 22 回大会,2005
8) 玉川奨,森田武史,日本語Ｗｉｋｉｐｅｄｉａからプロパティを備えたオントロジーの構築,人工知能学会論文誌,Vol26,No4,504-517,2011
9) 福本文代,辻井純一,コーパスに基づく動詞の多義解消,電子情報通信学会技術研究報告.NLC,言語理解とコミュニケーション,Vo94,No292,pp15-22,1994
10) Roberto Navigli,Universita di Roma La Sapienza,Rome,Itary,Word Sense Disambiguation: A Survey,ACM Computing Surveys,Vo41,No10,2009
11) Manuel de Buenaga Rodriguez,Jose Maria Gomez Hidalgo,Belen Diaz Agudo,Using Wordnet to complement Training Information in Text Categorization, In journal of Second International Conference on Recent Advances in Natural Language Processing, Vol.cmp-lg/9709007, 1997
12) Claudia Lecock,Educational Testing Serveice,George A. Miller,Princeton University,Martin Chodorow,Hunter Colleqe of CUNY,Using Corpus Statistics and WordNet Relations for Sense Identification, Computational Linguistics - Special issue on word sense disambiguation archive,Volume 24 Issue 1,pp147-165,1998
13) http://nlpwww.nict.go.jp/wn-ja/ (日本語 WordNet)
14) 池原悟,白井諭,横尾昭男,中岩浩巳,小倉健太郎,大山芳史,林良彦,日本語語彙体系,岩波書店,1997
15) http://wordnet.princeton.edu/ （WordNet）
16) http://www.illc.uva.nl/EuroWordNet/ (Euro WordNet)
17) http://www.tanos.co.uk/jlpt/
18) http://www.aozora.gr.jp/
19) http://sherlock-holm.es/
20) http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html(MeCab)