静岡大学　博士論文

ペアワイズアライメントを用いた動詞の多義性解消に関する研究

平成17年2月

静岡大学大学院理工学研究科
設計科学専攻

山下 浩一

目次

1 序論 **1**

1.1 研究の背景と目的 · · · · · 1

1.2 本論文の構成 · · · · · 4

2 関連研究と本研究の位置付け **6**

2.1 自然言語の解析における曖昧性 · · · · · 6

2.1.1 形態素解析 · · · · · 7

2.1.2 構文解析 · · · · · 8

2.1.3 意味解析 · · · · · 11

2.2 多義性解消の概略 · · · · · 13

2.2.1 単語の多義性の定義 · · · · · 14

2.2.2 多義性の分類 · · · · · 15

2.2.3 多義性解消の方針 · · · · · 16

2.3 先行研究 · · · · · 18

2.3.1 語義知識による分類 · · · · · 19

2.3.1.1 連想関係に基づく手法 · · · · · 19

2.3.1.2 選択制限に基づく手法 · · · · · 22

2.3.2 学習方法による分類 · · · · · 25

2.3.2.1 教師付き学習に基づく手法 · · · · · 25

2.3.2.2 教師なし学習に基づく手法 · · · · · 27

2.4 本研究の位置付け · · · · · 30

3 ペアワイズアライメントを用いた動詞の多義性解消 **32**

3.1 はじめに · · · · · 32

3.2 準備 · · · · · 34

3.2.1 単語の配列 · · · · · 34

3.2.2 ペアワイズアライメント ・・・・・・・・・・・・・・・・・・・・・・ 35
3.3 基本的な考え方 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 37
3.4 提案する手法 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 39
3.4.1 配列パターン ・・・・・・・・・・・・・・・・・・・・・・・・・・ 39
3.4.2 文脈の類似度の算出 ・・・・・・・・・・・・・・・・・・・・・・ 42
3.4.3 手法の適用例 ・・・・・・・・・・・・・・・・・・・・・・・・・・ 44
3.5 実験 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 46
3.6 考察 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 49
3.7 まとめ ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 51

4 アライメントスコアの重みの推定 53
4.1 はじめに ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 53
4.2 基本的な考え方 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 54
4.3 閾値の推定 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 55
4.4 重みの推定 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 58
4.5 実験 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 59
4.6 まとめ ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 62

5 文照合への応用 63
5.1 はじめに ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 63
5.2 文照合に関連する先行研究 ・・・・・・・・・・・・・・・・・・・・・・ 64
5.3 多義性解消の照合問題への応用 ・・・・・・・・・・・・・・・・・・・・ 66
5.4 実験 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 67
5.4.1 正解文の検索 ・・・・・・・・・・・・・・・・・・・・・・・・・・ 68
5.4.2 類似文の組の抽出 ・・・・・・・・・・・・・・・・・・・・・・・・ 70
5.5 まとめ ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・ 72

6 結論 73

謝辞 77

A 依存構造木へのノード “SUB”、“OBJ” の追加 84

B ペアワイズアライメントの導出 86

学位論文要旨

本論文は自然言語が持つ曖昧性の一つである単語の多義性のうち、動詞を対象とした多義性解消についての新しい試みをまとめたものである。多義性解消は、その代表的な応用として機械翻訳における訳語選択や情報検索における検索対象絞り込みなどが挙げられ、自然言語処理システムにおける有用性が極めて高い。このため、自然言語処理の最も初期の段階から多義性解消の問題は広く認識され、さまざまなアプローチによる研究が盛んに行われている。しかし、こうした研究の誕生から50年以上経った現在でも、単語の多義性の問題は充分に解決できているとは言えない。計算機とインターネットの急速な普及に伴なって、現在人間が相互に伝達している情報の量と多様性はこれまでにない速度で増大しつつある。こうした背景の下、計算機による自然言語情報の効率的な処理が強く求められており、高精度で高品質な自然言語処理システムが望まれている。本研究はこの要求に応えることを目標に行われたものである。

これまでの多義性解消の手法は、多義語に対する構文的な制約を手がかりにするものと、多義語の近傍に出現する単語の分布を手がかりにするものとに大別できる。これらの手法は用いる手がかりによって特徴づけられるが、多義性解消に対する手がかりの基本的な役割は、前者の手法では制約であり、後者の手法では選好である。すなわち、両手法で用いられている語義選択の手がかりは対照的な観点に基づいている。このため、両者を組み合わせた情報を用いて多義性解消を試みている研究は極めて少ない。しかし実際の自然言語には、構文的な制約の観点で多義性解消が不可能である事例と、単語の分布の観点で不可能な事例とが混在して出現する。従来の手法はこれらの手がかりのどちらか一方を他方とは独立に用いており、従って精度の向上には限界が考えられる。

本研究ではこの問題に対し、構文的な制約と近傍の単語の分布とを組み合わせた新しい手がかりを多義性解消に用いる手法を構築した。本手法は従来の二つの手法の特長を併せ持つものであり、これによって従来より高い精度での多義性解消を可能としている。計算機による多義性解消では、語義ごとに与えられた知識と多義語の文脈から得られた手がかりとの類似性を判断する必要があるが、本手法はこの判断にペアワイズアライメントの技法を用いる。これによって本手法は、言語の効率性が高い自然言語を対象に柔軟で頑

健な処理を行うことが可能である。本論文では本研究で構築したこの新しい手法について詳述し、評価実験を通じて本手法を用いた動詞の多義性解消が平均で81.1%の精度を達成したことを示す。

本研究で構築した多義性解消の手法は、実装のコストの観点でいくつかの問題を有する。特に、語義に関する知識獲得や入力文の構文解析などの人手による調整が介在しており、このときの人手のコストが実装コストの多くを占める。本研究で構築した多義性解消の手法を応用するものとしては機械翻訳システムや情報検索システムなどがある。こうしたシステムに本手法を応用するためには、あらかじめ語義知識をシステムの辞書情報として与えておく必要がある。本手法の実装には、このときの人手のコストが大きな障壁となると考えられる。この問題に対し、本論文では人手による知識獲得のコストを軽減させる試みについて述べる。具体的には、大規模コーパスからの統計情報を利用して語義知識の一部を推定する手法について説明する。統計情報を利用することによって獲得する知識の品質は若干低下するものの、複数の動詞に対して従来の手法よりも高い精度で多義性を解消できることを示す。

また、本研究で構築した多義性解消の手法を文の照合の問題に応用することについて検討する。文照合のさまざまな応用において、表層文字列の水準での照合では、要求される類似性判断の精度を満足する結果が得られない場面が多数存在する。照合に表層文字列だけでなく文の構文情報まで用いることは、文照合の精度を向上させるための妥当な展開であると考察される。このとき、構文情報の類似性をいかにして評価するかが問題となる。こうした観点の下、本論文では多義性解消と情報検索の高い関連性に着目し、本研究で構築した多義性解消の手法を文照合の問題に応用することについて検討した。本応用の妥当性を検証するための二つの実験から期待された結果が得られ、本応用が有望であることが示唆された。本論文ではこの応用と実験についても詳述し、手法の妥当性を別の観点から明らかにする。

以上、本研究で構築された多義性解消の手法は、実装のコストの観点でいくつかの問題を有するものの、従来の手法よりも高い精度で多義性を解消することが可能である。本論文では、本研究で構築した多義性解消の手法の詳述、実装コスト軽減のための試みの報告、多義性解消以外の問題への応用の検討と、大きく三つについて論じた。本論文はこれらの論述を通じて、高精度・高品質の自然言語処理システム構築のための一手法を示すものである。

第 1 章

序論

1.1 研究の背景と目的

言語(language)とは音声や文字によって任意の情報を表現・伝達・理解するための規則や体系のことである。一般に言語は人工言語(artificial language)と自然言語(natural language)に大別される。人工言語とは人間がある目的のために設計した言語である。代表的な人工言語の例には数学的記法の体系やプログラミング言語などのように特化した目的のために設計されたものが多いが、例外的なものとして人間同士のコミュニケーションのためのエスペラント語やノシロ語などがある。

一方、自然言語とは日本語や英語など、人間が意思の疎通や情報の伝達などに日常的に用いている言葉のことである。意思疎通や情報伝達は、人間が社会を形成して社会的生活を営む上で必要不可欠な行為であり、人間はこの要求を満足するものとして言語を自然発生的に生み出した。自然言語の代替には身ぶり手ぶりや顔の表情などがあるが、言語的に体系化された一部の例外を除けば、これらの手段の表現能力の貧困さは否定できない。自然言語は人間にとって最も自然で高度な情報伝達の手段を提供するものと言える[1]。

現在、人間と言語を取り巻く環境は大きく変化している。計算機とインターネットは急速に普及し、人間相互の情報伝達において物理的な距離の影響は小さくなりつつある。また、これに伴って日々膨大な情報が人々の間で交換されるようになっている。こうした状況の下、計算機によって自然言語を処理するシステムの高性能化・高品質化が強く求められており、その基盤となる自然言語処理研究のさらなる発展が望まれている。

自然言語処理(natural language processing)とは、計算機による自然言語の

さまざまな処理を扱う研究分野である。自然言語処理の研究の歴史は古く、Warren Weaverが1947年にマサチューセッツ工科大学のNorbert Wienerに宛てて書いた手紙がその始まりとされるのが一般的である[2]。Weaverの手紙は計算機による翻訳、すなわち機械翻訳(MT; machine translation)の実現可能性について書かれたものであった。それから50年以上にわたり、自然言語処理の研究は機械翻訳の研究を中心としてさまざまに発展してきた。現在では、商用化された機械翻訳システムも数多く登場している。しかし、現在のシステムでは翻訳結果に充分な精度が達成できておらず、自然言語処理の研究は未だ発展途上にあると言える。

自然言語処理を困難にしている要因の中で、最も主要なものの一つに曖昧性(ambiguity)の問題がある。自然言語の曖昧性とは、一つの表現が複数の異なった解釈を持つ性質をいい、自然言語処理で扱われる曖昧性の問題とは、計算機を用いた自然言語解析において複数の解析結果が解として許される問題のことである。曖昧性の問題の重要さや困難さは、自然言語処理の最も初期の段階から認識されてきた。事実WeaverはWienerへの手紙の中で、複数の解釈が存在することによる意味的な困難さが機械翻訳の実現を否定する可能性を示唆している。

人工言語は言語設計者によって語彙や文法が人工的に作成されるため、表現とそれに含まれる情報とが一対一に対応し、曖昧性の問題が存在しないという共通した性質を持つ。一方、自然言語は同じ表現が文脈に依存して異なる情報に対応するため、人工言語と比較して言語の効率性(efficiency of language)は極めて高い。自然言語が曖昧性の問題を持つのは、こうした言語の高い効率性の代償である[3]。自然言語処理システムでは曖昧性の扱いがシステムの品質に直接結びつくため、曖昧性解消は自然言語処理研究に課せられた主要な課題と言える。

本論文は自然言語が持つ曖昧性のうち単語の多義性(polysemy)の問題に焦点を絞り、動詞を対象とした多義性解消(WSD; word sense disambiguation)についての新しい試みをまとめたものである。一般に複数の意味を持つ単語は多義語(polysemous word)と呼ばれる。例えば“bank”は「土手」の意味と「銀行」の意味を持つ多義語である。この性質により、“Sitting on the bank, I was looking at the river.”という文(sentence)には次のように二通りの解釈が存在することになる。

1. 土手に腰を降ろして、私は川を眺めていた。

2. 銀行に腰を降ろして、私は川を眺めていた。

この場合には1.の解釈を選択するよう、“bank” の多義性を解消しなければならない。こうした処理を行うのが多義性解消の役割である。

本研究で多義性解消を取り扱う最も大きな動機付けは、その有用性にある。多義性解消の最も主要な応用は機械翻訳システムである。すなわち、文脈(context)に依存して “bank” の日本語表現を「土手」とするか「銀行」とするかを決定するように、多義性解消を訳語選択に応用するものである。日英機械翻訳システムを対象とした麻野間らの調査[4]によると、機械翻訳の精度を低下させている要因のうち約40%は、適切な訳語が選択できないことにある。従って多義性解消の精度向上が、機械翻訳における翻訳精度の大幅な向上に繋がると期待できる。

また、多義性解消の別の応用として情報検索(information retrieval)システムが挙げられる。検索対象の文書に含まれる単語が多義であるとき、その語義を正しく決定することは検索精度の向上に寄与する。例えば “Java” という単語がプログラミング言語を意味するのか、コーヒーを意味するのか、ジャワを意味するのかが明らかであれば、検索精度が向上するであろうことは容易に予想できる。Schützeらは、多義性解消によって情報検索システムの検索精度が7%から14%まで向上することを報告している[5]。

このように、自然言語処理システムにおける多義性解消の役割りは極めて大きく、多義性解消は自然言語処理における最も基礎的かつ重要な課題の一つとして位置づけられる。しかし、現在の機械翻訳システムの翻訳精度や、情報検索システムの検索精度は充分な精度が達成できているとは言えない。こうした背景の下、本研究では自然言語処理システムの高品質化を目標として多義性解消の精度を向上させることを目的とする。

これまでの多義性解消の手法は、多義語に対する構文的な制約を手がかりにするものと、多義語の近傍に出現する単語の分布を手がかりにするものとに大別できる。これらの手法は用いる手がかりによって特徴づけられるが、多義性解消に対する手がかりの基本的な役割は、前者の手法では制約であり、後者の手法では選好である。すなわち、両手法で用いられている語義選択の手がかりは対照的な観点に基づいている。このため、両者を組み合わせた情報を用いて多義性解消を試みている研究は極めて少ない。しかし実際の自然言語には、構文的な制約の観点で多義性解消が不可能である事例と、単語の分布の観点で不可能な事例とが混在して出現する。従来の手法はこれらの手が

かりのどちらか一方を他方とは独立に用いており、従って精度の向上には限界が考えられる。

本研究で構築した手法はこの問題に対し、構文的な制約と近傍の単語の分布とを組み合わせた新しい手がかりを多義性解消に用いるものである。すなわち本手法は従来の二つの手法の特長を併せ持つものであり、これによって従来より高い精度での多義性解消を可能としている。計算機による単語の多義性解消では、語義ごとに与えられた知識と多義語の文脈から得られた手がかりとの類似性を判断する必要があるが、本手法はこの判断にペアワイズアライメントの技法を用いる。これによって本手法は言語の効率性の高い自然言語を対象に、柔軟で頑健な処理を行うことが可能である。本論文は本研究で構築したこの新しい手法についてまとめたものであり、複数の評価実験を通じて本手法の妥当性について論じるものである。

1.2 本論文の構成

本論文の構成は以下の通りである。

第1章では本研究の概要と動機付けを明確にする目的で、自然言語処理研究の歴史と本研究の背景とを概説する。本研究で扱う多義性解消の有用性が極めて高いことを示し、多義性解消の精度を向上させることが自然言語処理システムの高精度化・高品質化に繋がることを示す。また、従来の手法が持つ問題点について概要を示し、本研究で構築した手法の大まかな特徴を述べる。最後に本節を通じて本論文の構成を明らかにする。

第2章では、本論文で扱う多義性の問題の位置付けを明確にする目的で、自然言語処理における曖昧性の問題を概説する。また、本研究の位置づけを明確にするために、語義と多義の定義について言及するとともに単語の多義性解消のためにどのような研究がなされてきたかについて述べる。これらの先行研究について複数の観点による分類を示し、本研究の位置づけを示す。これによって、従来の手法と本研究で構築した手法との関係を明確にする。

第3章では本研究で構築した多義性解消の手法について述べ、この手法が動詞に対して高い精度で多義性解消可能であることを示す。これまでの多義性解消の手法では正しい解が得られない事例があることを示し、従来の手法では精度の向上に限界があることを示す。この問題に対して本研究で構築した手法について詳説し、本手法が従来の手法では正解を導けなかった例に対し

て正しい処理を行えることを示す。また、評価実験を通して、本手法が動詞に対して平均81.1%の精度で多義性解消可能であることを示す。

第4章では、第3章で述べる手法において問題となる手法の実装のコストを軽減させるために、語義知識獲得のコストを軽減する試みについて述べる。具体的には、大規模コーパスからの統計情報を利用して語義知識の一部を獲得する手法について示す。統計情報を利用することによって獲得する知識の品質は若干低下するものの、複数の動詞に対して従来の手法よりも高い精度で多義性を解消できることを示す。

第5章では、第3章で述べる手法を文の照合の問題に応用することについて検討する。文の照合とはどのような問題であるのかについて明らかにし、この問題に関連する先行研究について概説する。また、文照合の問題に適用するために本研究で構築した手法に対して施した若干の変更について言及し、本応用における文照合を定式化する。実験結果の詳述を通して、文照合の観点からの本手法の妥当性について述べる。

第6章では本論文を通しての結論を述べる。各章での論述について総括し、本研究に積み残された課題について述べる。

以上、本論文では、動詞の多義性解消においては本手法が既存の手法と比較して精度の面で優れていること、コーパスからの統計情報を利用することによって本手法の実装のコストが軽減できること、多義性解消以外の問題への応用を通じて本手法の妥当性と有用性が高いことなどを述べる。

第 2 章

関連研究と本研究の位置付け

自然言語処理では、自然言語の解析の際に複数の解析結果が頻繁に得られる。この問題は曖昧性の問題と呼ばれ、複数の解析結果から最も妥当な解を選択することを曖昧性解消と呼ぶ。

本論文で扱う単語の多義性の問題は自然言語処理で扱われる曖昧性の問題の一つとして位置付けられる。本章では研究の位置付けを明確にする目的で、自然言語処理で扱われる曖昧性の問題について言及し、単語の多義性がこれらの曖昧性においてどのように位置付けられるのかを明らかにする。また、本研究の対象となる単語の多義性とはどのような問題であるのかについて説明する。さらに、本論文に関連するこれまでの研究事例を整理し、本研究がこれらの先行研究においてどのように位置付けられるのかを明確にする。

2.1 自然言語の解析における曖昧性

自然言語の解析システムは、自然言語の文法をモデル化した有限個の解析規則を用いて文の解析を行う。しかし、自然言語は無限とも言えるほどの多様な事柄を表現可能という性質を持っており、すべての文脈とすべての語彙 (lexicon) に対して固有の解析規則を用意することは極めて困難である。このため、通常解析規則は複数の事例を一般化したものとして開発される。規則の一般化は、機械による自然言語解析を現実的なものとする反面、判断情報の欠落などから複数の解析結果を産み出し、曖昧性の原因の一つともなる。

曖昧性の問題の現実的な解決策としては、解析結果に優先順位を付与することが一般的である。本節では以下、自然言語処理における代表的な解析技術として、形態素解析 (morphological analysis)、構文解析 (parsing)、意味解析 (semantic analysis) を取り上げ、それぞれの解析技術を簡単に説明し、そこで

生じる曖昧性の問題について概説する。

2.1.1 形態素解析

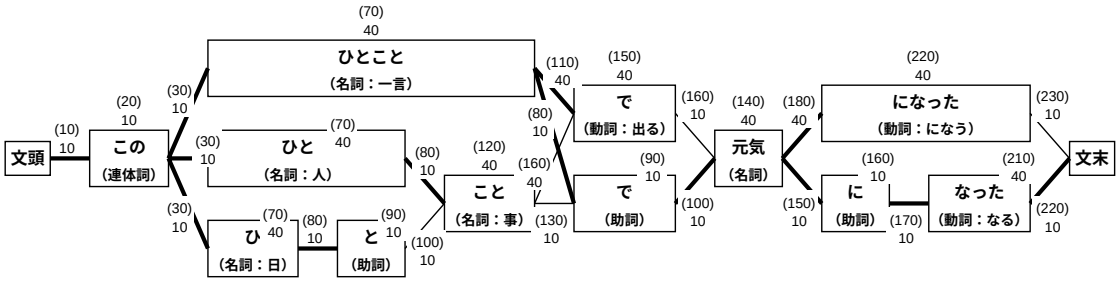
形態素解析とは言語の最小の意味の単位である形態素(morpheme)を入力文から抽出する解析技術である。形態素解析は、連続した文字列として入力される自然言語の文を対象に、入力文に含まれる形態素の同定(word segmentation)と、形態素への品詞情報の付与(part-of-speech tagging)との二つの処理を中心とする。ここで、英語などに代表される単語を区切って表記する言語では、形態素の同定は単純であるため、主として品詞の割り当てのみが中心に扱われる。一方、日本語などに代表されるいわゆるべた書き文で表記する言語では、形態素の同定が必要となるため、形態素の同定と品詞割り当てが同時に行われることが多い。

形態素の同定に関しては特に日本語の解析の際に曖昧性が問題となる。例えば「今日本人が会社に来た」という文は「今日/本人-が/会社-に/来-た」という形態素列と、「今/日本人-が/会社-に/来-た」という形態素列の二通りの解釈が存在し、曖昧さを持つ。一方、品詞の割り当てに関しては英語でも日本語でも曖昧性が問題となる。例えば“box”という単語は名詞と動詞の二通りの品詞が付与される可能性があり、曖昧さを持つ。これらの解析結果に対しては最長一致法や文節数最小法などのヒューリスティクスを用いた優先度付与や、最小コスト法や最尤法などを用いて品詞付与と同時に優先度を付与することで曖昧性を解消する手法が知られている。

現在最も広く用いられている日本語形態素解析システムの一つにJUMAN[6]があるが、ここではJUMANに採用されている最小コスト法について簡単に説明する。最小コスト法ではまず、

- 辞書を参照して入力文中の各位置から始まる単語を取り出し、
- 単語と単語の連接可能性をチェックしながら取り出された単語をつないでいく

という二つの処理を繰り返し行うことによって、単語をノードとするラティス構造(lattice structure)を生成する[7]。ここで、この二つの処理を実行する際には、単語の品詞、読み、活用形などを規定する単語辞書と、行列の形式で連接可能な二語を規定する連接可能性辞書とが必要となる。連接可能性とは、二つの単語が連続して文中に出現する可能性を意味している。次にラティス解



(括弧内の数値は各ノード/リンクまでの部分最小コストを、括弧外の数値は各ノード/リンクに与えられたコストを示し、太線のリンクは部分最適解を示す)

図 2.1: 最小コスト法による形態素解析の例[7]

のノードとリンクに単語とその連接の重みに準ずるコストを与え、コスト最小の経路(ノードとリンクの並び)を優先解として選択する。例えば図2.1は、「このひとことで元気になった」という入力文に対して最小コスト法による形態素解析を行った結果を示すものである。

最小コスト法による形態素解析では、形態素の連結の規則を二語の間に存在する連接可能性に一般化する。また、品詞選択の規則に関しては各単語が独立に品詞と対応する規則に一般化する。これらはそれぞれ連接可能性辞書と単語辞書の参照に相当し、こうした一般化に伴なって解析結果は複数得られることになる。最小コスト法は単語の品詞選択と、二語の連接にコストを与えることによって解析結果に優先順位を付与し、曖昧性の解消を図っている。

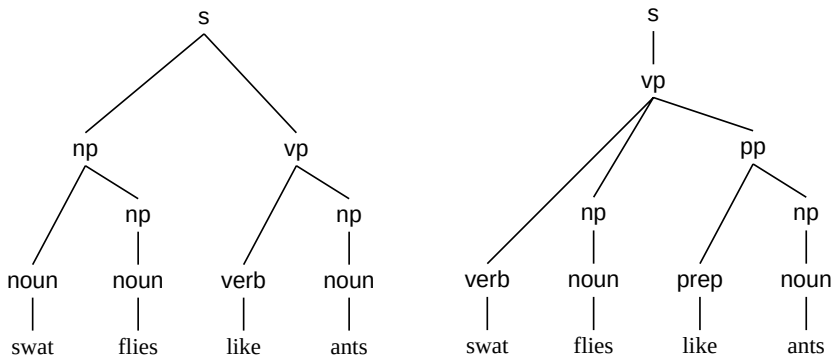
2.1.2 構文解析

構文解析とは、文の文法的な構造である構文構造(syntactic structure)を明らかにするための解析技術である。文中の単語間には修飾関係が存在し、修飾関係の連鎖によって文は一つの構文構造を持つ。しかし、文は表記・発話される時点で一次元の単語の並びに変換される。すなわち構文解析は書き手や話し手が意図した文の構造を復元する処理と換言できる。

構文解析で中心となる処理は、構文的整合性に照らして入力文の構文構造の候補を探索することである。構文的整合性を表すものとしては、文法的知識や単語の用法に関する知識が用いられる。これらの知識に関しては、文法や解析アルゴリズムによってさまざまな形式が存在する。ここでは例として、英語文の構文解析に広く用いられている文脈自由文法(context free grammar)

s → **np vp**	**vp** → **verb**	**prep** → like	**noun** → flies
s → **vp**	**vp** → **verb np**	**verb** → swat	**noun** → ants
np → **noun**	**vp** → **verb pp**	**verb** → flies	
np → **noun pp**	**vp** → **verb np pp**	**verb** → like	
np → **noun np**	**pp** → **prep np**	**noun** → swat	

図 2.2: 文脈自由文法[8]



解析木t_1　　解析木t_2

図 2.3: 構文解析の結果の一部[8]

を取り上げる。

文脈自由文法Gは四つ組$\langle N, \Sigma, P, S\rangle$で定義される。四つ組の記号$N$は非終端記号(non-terminal)の集合、$\Sigma$は終端記号(terminal)の集合、$P$は書き換え規則(production rule)の集合、及びSは出発記号(start symbol)を表す。文脈自由文法では、Pの要素は$A \to \alpha$ $(A \in N, \alpha \in (N \cup \Sigma)^*)$の形式を満たす規則に限られる。今、図2.2に示す文法が与えられたとすると、文“Swat flies like ants”は図2.3のように複数の解析結果が存在し、曖昧さを持つ。ここで、代表的な解析アルゴリズムとしてはCKY法、チャート法、アーリー法、一般化LR法などが知られているが、詳細は割愛する。

構文解析で生じる曖昧性に関してはこれまでに多くの研究がなされている。文脈自由文法を用いた解析では、構文的優先度に関するヒューリスティクスや、選択制限(selectional restriction)に基づく意味的制約を取り入れて解析候補を絞り込む試みなどが報告されている。また、文脈自由文法を拡張した確率的文脈自由文法(PCFG; probabilistic CFG)を用いて、書き換え規則に対する

s → **np vp**	: 0.8	**prep** → like	: 1.0
s → **vp**	: 0.2	**verb** → swat	: 0.2
np → **noun**	: 0.4	**verb** → flies	: 0.4
np → **noun pp**	: 0.4	**verb** → like	: 0.4
np → **noun np**	: 0.2	**noun** → swat	: 0.05
vp → **verb**	: 0.3	**noun** → flies	: 0.45
vp → **verb np**	: 0.3	**noun** → ants	: 0.5
vp → **verb pp**	: 0.2		
vp → **verb np pp**	: 0.2		
pp → **prep np**	: 1.0		

図 2.4: 確率的文脈自由文法[8]

選好を取り入れる試みや、単一化文法(unification grammar)、主辞駆動句構造文法(HPSG; head-driven phrase structure grammar)などを用いて文法規則の上での制約を精密にし、適用可能な書き換え規則を絞り込む試みなどが報告されている。ここでは確率的文脈自由文法による解析結果の優先順位付けを取り上げて概説する。

確率的文脈自由文法G'は四つ組$\langle N, \Sigma, P', S\rangle$で定められる。文脈自由文法との違いは、任意の$A \in N$に対して生成規則$A \to \alpha$が確率$\Pr(A \to \alpha)$を持つ点である。ここで、各$A$生成規則に対して$\sum_\alpha \Pr(A \to \alpha) = 1$が成り立つ必要がある。すなわち、同じ左辺を持つ規則の確率は合計すると1になるという条件である。今、単語列$w_1, w_2, \cdots, w_n$からなる入力文に対し、出発記号Sから導出木tを生成する導出を$S \overset{p_1}{\Longrightarrow} \alpha_1 \overset{p_2}{\Longrightarrow} \alpha_2 \overset{p_3}{\Longrightarrow} \cdots \overset{p_m}{\Longrightarrow} \alpha_m = w_1, \cdots, w_n$とすると、$t$を生成する確率は$\Pr(t) = \prod_{i=1}^{m} \Pr(p_i)$で定義される。今、図2.4に示す確率的文脈自由文法が与えられたとすると、文“Swat flies like ants”の解析結果t_1、t_2には、$\Pr(t_1) = 3.456 \cdot 10^{-5}$、$\Pr(t_2) = 2.88 \cdot 10^{-4}$のように確率的な優先順位を付けることができる。

文脈自由文法に基づく構文解析では、書き換え規則の性質から文脈に依存しない、すなわち文脈情報を考慮しないという一般化がなされている。また、通常用いられる書き換え規則の上では、修飾関係が品詞のレベルにまで一般化されている。これらの一般化の下で書き換え規則の数は制限され、解析アルゴリズムを単純なものにできる反面、複数の木の導出を招く。この曖昧性

を解消するために、確率的文脈自由文法では確率による選好が導入され、導出木に優先順位がつけられる。

2.1.3 意味解析

意味解析とは文の意味構造を明らかにするための解析技術である。意味構造とは文が伝える意味を表現するもので、誰が、何を、誰に、いつ、どこで、どのように、何をしたか、などの情報を明示的に含むものである。意味解析ではこれらの情報を取り出すために、文中の単語の語義や単語と単語の間の意味関係などを解析する。すなわち、本論文で扱う単語の多義性に関する解析技術は意味解析の一部として捉えられることが多い。単語の多義性の問題や多義性解消の概略などは2.2節で言及することとし、ここでは単語間の意味関係の解析について概説する。

一般に、文中で依存関係にある単語間の意味関係は、表層表現から必ずしも一意に決まらない。例えば、日本語の助詞「の」で結ばれる二つの名詞の間の意味関係を取り上げる。「の」で結ばれる二つの名詞とは、「AのB」という名詞句を構成するものである。こうした名詞句の出現頻度は高く、かつその意味内容は極めて多様である。「AのB」に関する詳細な分析としてよく知られたものに島津らの報告[9]があるが、島津らは「AのB」の意味関係を表2.1のように86種に分類している。例えば「彼のメガネ」における所有関係、「太郎の結婚」における動作主-述語関係、「航海の技術」における述語-対象関係など、品詞レベルの表層表現には多数の意味関係が考えられ、曖昧さを持つ。

島津らは「AのB」の意味関係の曖昧性解消として、詳細な意味的制約を用いる手法を報告している。ここでの意味的制約とは、二つの名詞AとBに関する制約であり、この制約を記述するために素性(feature)と呼ばれる概念が導入される。素性とは単語の属性や機能を表すもので、単語の主要な意味特徴を表す主素性(「椅子」に対してはthing、「犬」に対してはanimate、「遊び」に対してはactionなど)、他の素性との意味的依存関係を表す依存素性(「日本人」に対しては[belong-to nation]、「遊び」に対しては[agent animate]など)、他の単語との結合の仕方や結合における役割を表す機能素性(「公園」に対しては[pos noun]、「人間」に対しては[role agent]など)の三種類が設定される。通常、各単語にはこれらの素性が複数割り当てられる。

島津らの意味関係解析では、AとBの素性のうち、あらかじめ与えられた制

表 2.1: 朝日新聞・天声人語における意味関係の出現頻度 [9]

意味関係	出現数	意味関係	出現数	意味関係	出現数
動作主-述語	326	状態・様態の指定	69	述語-様態	22
対象-述語	401	結果の指定	26	述語-程度	18
随伴対象-述語	21	対象の指定	126	述語-数量	4
与え手-述語	8	大きさの指定	22	所有関係	426
受け手-述語	14	色等の指定	35	人間関係	44
手段-述語	12	温度等の指定	4	全体・部分	230
道具-述語	2	形・構造の指定	72	部分・全体	8
材料-述語	3	機能・性能の指定	61	数量で限定	246
原因-述語	7	性質・属性の指定	27	年齢で限定	40
時-述語	115	名称の指定	26	順序で限定	30
場所-述語	81	数量の指定1	53	種類で限定	99
起点-述語	17	数量の指定2	14	役割で限定	56
着点-述語	43	数量の指定3	40	程度で限定	125
目的-述語	23	役割・目的の指定	17	性状で限定	237
場合-述語	13	述語-動作主	38	材料で限定	36
内容-述語	48	述語-対象	83	原因で限定	44
様態-述語	53	述語-随伴対象	1	作者で限定	95
回数-述語	2	述語-手段	22	生産物で限定	2
割合-述語	6	述語-道具	14	場所で限定	401
程度-述語	15	述語-材料	3	所属で限定	256
数量-述語	33	述語-原因	14	時で限定	324
順序-述語	4	述語-場所	16	起点で限定	83
場所の指定	148	述語-時	16	着点で限定	41
時の指定	102	述語-起点	3	状況で限定	82
範囲の指定	110	述語-着点	4	目的で限定	93
方向の指定	11	述語-方向	8	内容で限定	233
目的の指定	47	述語-回数	0	指示で限定	57
原因の指定	18	述語-目的	3	特定化で限定	40
状況の指定	68	述語-割合	1		

約の下で親和的なものが結合して意味関係が求められる。例えば「子供の遊び」では「子供」のanimateという素性と「遊び」のactionという素性が、「生物は動作する」及び「その生物は動作主だ」という制約の下で結びついて、動作主-動作という意味関係を決めることができる。このように素性からの意味関係の導出を行う演算としては、素性ユニフィケーションと呼ばれる単一化(unification)の概念を利用した演算が用いられる。素性ユニフィケーションの最も基本的なものは、Prologの記法を用いて具体的に次のように表すことができる。

```
rel-unify(A, B, R) :- member([R, X], B), member(X, A)
```

ここで、Aは主素性のリスト、Bは依存素性のリストを表す。また、rel-unify(A, B, R)はAとBとの意味関係がRであることを表し、member(X, Y)はXがリストYの要素であることを表す。このとき、例えば名詞句「さるの食事」の意味関係を導出するゴール(goal)は「さる」の主素性と「食事」の依存素性を用いて次のように表現される。

```
rel-unify([animate], [action, [agent, animate]], R)
```

すなわち、「さるの食事」の意味関係は単一化によって求められるRへの代入である。この例ではR = agentと正しい意味関係を求めることができる。

「AのB」の意味関係解析では、品詞のような表層的な一般化では意味関係の細かな差異を表現することができず、多数の解釈を許すことになる。素性ユニフィケーションによる意味関係解析では、単語の機能的・意味的特徴の差異が表現できる程度に単語の一般化を抑制し、素性を用いた表現を導入している。これによってAとBの間の意味的制約を素性のレベルの粒度にまで細分化して規則化することができる。但し、同一の素性を持つ単語間の機能的・意味的特徴の差異までは表現できないため、単語に付与する素性の粒度をどの程度に設定するかについて、充分な検討が必要となる。

2.2 多義性解消の概略

多義性解消とは、多義語の適切な語義を文脈から同定することである。本節では多義性解消の概略を明らかにすることを目的に、本研究で対象とする単語の多義性について概説し、これを解消するための枠組みを概説する。こ

れらの概説の前に、多義語の概念や語義の概念を明確に定義することによって多義性解消の問題を明確にする。

2.2.1 単語の多義性の定義

語義、すなわち単語の意味とは何であるのかという問題に対しては、これまでにさまざまな議論が重ねられてきている[1, 10, ほか]。しかし、この問に対する普遍的で厳密な答えは存在しない。これは人間が持つ意味に関する心的表現が未だほとんど明らかにされていないためである。意味の心的表現は心理学的実験によって明らかにすることが期待されるが、このような実験の設計は極めて困難と言える。このため、語義にはさまざまな定義があり、例えば単純に辞書項目をそのまま語義として利用する定義や、語義はほかの語句との関連で生じるもので単独では存在しないという定義もある。

本研究では語義に辞書項目を利用する定義を採択する。すなわち、任意の単語wの語義はwのみに依存して存在し、$\{s_1, s_2, \cdots, s_n\}$のように集合で表すことができる。$s_1, \cdots, s_n$は$w$の辞書における定義項目と1対1に対応する任意の記号である。今、wの語義(辞書における定義項目)の集合を$SENSE(w)$で表し、$SENSE(w)$の要素の個数を$|SENSE(w)|$で表すとすると、多義語とは$|SENSE(w)| \geq 2$を満たすwと定義することができる。

多義語によって解釈に曖昧性が生じる例として最も有名なものの一つに、Bar-Hillelの指摘がある[11]。Bar-Hillelは計算機による多義性解消の困難さを示すために、次の文章を例示した。

> Little John was looking for his toy box.
> Finally he found it.
> The box was in the pen.
> John was very happy.

ここで、“pen”は「(筆記用具としての)ペン」の意味と「囲い」の意味を持つ多義語である。従ってこの文章には二通りの解釈が存在することになり、曖昧さを持つ。多義性解消とは、入力における多義語の文脈から、多義語の適切な語義を選択することである。すなわちこの例において多義性解消とは、上に示した文章を用いて多義語“pen”の意味を「ペン」と「囲い」のどちらかに決定することである。

2.2.2 多義性の分類

語義を辞書項目として定義した場合、単語の多義性は複数の組に分類することができる。ここではWeissによる多義性の分類[12]について概説し、本研究で対象とする単語の多義性が、これらの分類の中でどのように位置付けられるのかを明らかにする。Weissによる分類では、単語の多義性は次のように三つに分けられる。

True Ambiguities 一つの単語が二つ以上の意味的機能(semantic function)を持つような曖昧性をいう。意味的機能とは、単語がその文脈に及ぼす影響のことである。例えば単語“bottom”は“of the bottle”という文脈に適用すると「ボトルの下部」という意味を生成し、“of the inning”に適用すると「(野球における何回かの)裏」という意味を生成する。こうした影響を意味的機能と呼ぶ。True Ambiguitiesのこのほかの例としては“degree”が計測単位や学位などを意味することが挙げられる。

Contextual Ambiguities 意味的機能が一つしか存在しない単語が複数の語義を持つような曖昧性をいう。例えば次の例における“base”を考える。

- first base (baseball)
- military base
- lamp base
- base register

これらの“base”は表面的に多義であるが、意味的機能は一つである。“base”はそれぞれ各文脈(“first”や“military”など)の基本的な側面、あるいは最も重要な側面を表している。

Syntactic Ambiguities 語義が品詞に依存して決定されるような曖昧性をいう。例えば“Sam plays in the park.”と“Sam likes the play.”における“play”は品詞が異なっており、このことから語義は全く異なるものになっている。

多義語はその文中における出現に応じて、これらの三つのクラスの一つ、もしくは複数に対応付けられる。本研究で曖昧性の解消の対象とするのは、このうちのTrue AmbiguitiesとContextual Ambiguitiesである。

Syntactic Ambiguitiesに関しては、形態素解析や構文解析で用いられる文法的な知識を用いることによって曖昧性が解消可能であり、従って本論文では対

象から外すものとする。一方、文法的な知識だけを用いて多義性が解消できないという点でTrue Ambiguitiesと Contextual Ambiguitiesとの明確な差異は存在しない。また、本研究では2.2.1節で言及したように語義を集合として定義しており、語義の集合が機械可読辞書などの外部情報からあらかじめ獲得できることを想定している。従って本研究で解消すべき多義性は外部情報における定義に依存し、True Ambiguitiesと Contextual Ambiguitiesとの扱いを区別することは不可能である。このため、本論文ではSyntactic Ambiguitiesは処理の対象とせず、True Ambiguitiesと Contextual Ambiguitiesを多義性解消の処理の対象と見なす。

2.2.3 多義性解消の方針

ここで、本研究で扱う多義性解消の問題を整理する。本研究で想定する多義語wの語義はwのみに依存して定義される。今、wの語義を$\{s_{11}, s_{12}, \cdots, s_{mn}\}$で表すものとする。ここで、語義$s_{ij}$の$i$は$w$の品詞を表す数であり、$j$は$w$の品詞$i$における語義を表す数である。本研究でいう多義性解消とは、新しく入力された多義語wに対し、あらかじめwに対して定義された語義の集合から妥当な語義を選択することである。語義の選択には、入力された多義語wの文脈を手がかりに用いる。このとき、wに対しての形態素解析はすでに終了しているものと仮定し、wは品詞情報を伴なって入力されるものとする。すなわち、入力されたwの品詞がiであったときには、語義選択の範囲は$\{s_{i1}, \cdots, s_{ik}\}$に限定される。本節では以下、この問題を解決するための方針について、その概要を示す。

Weissは2.2.2節に挙げた文献[12]の中で、人間が自然言語における単語の多義性を解決するために、大きく二つの情報源を利用していることを指摘している。

- 文脈に含まれる手がかり
- 読者の持つ実世界の情報の蓄積

しかし、これらの情報を用いた人間による多義性解消という認知的な処理をモデル化することは極めて困難である。特に、読者の持つ実世界の情報の蓄積は常識(common sense)としてしばしば参照される概念であるが、この知識体系は極めて複雑である。人間の常識を広範囲にわたって収集・蓄積する試みは、一部の例外(Cyc Projectなど)を除いてほとんどなされていない。

多義性解消に用いられる文脈についての分析としては、Kaplanの報告[13]がよく知られている。Kaplanは七人の被験者を対象にして、人手による語義の曖昧性解消にはどの程度の範囲の文脈が必要かを明らかにするための実験を行っている。Kaplanは多義語に対し、左隣の単語(P1)、右隣の単語(F1)、P1とF1の両方(B1)、左隣の二単語(P2)、右隣の二単語(F2)、P2とF2の両方(B2)、文全体(S1)の七種類の文脈を被験者に与え、多義性解消を行わせた。この結果から、被験者が正しく語義を選択した割合を元として、各文脈が多義性を減少させる割合を求めている。これによると、S1は平均して多義性を26%にまで縮小させ、一方、B1,B2はそれぞれ33%, 36%にまで縮小させている。このことからKaplanは、多義語の左右に隣接する二単語、あるいは左右二単語ずつの四単語から構成される文脈は、多義性解消の手がかりとして文全体から構成される文脈と同程度の効果があるという結論を導いている。

しかし、Kaplanの実験における七人の被験者は、多義語の左右に隣接する僅かな単語以外の情報を用いなかったのではなく、実世界における膨大な量の情報の蓄積を活用して多義性を解消したものと考えられる。Weissによって指摘された二つの情報源は互いに強く依存し合うものであり、独立して用いられるものではない。Kaplanの心理学的見地からの所見に対し、Galeらは次のように反論している[14]。

> — However, as has been found in chess playing programs, attempting to model the way people do things may not be the best way for a computer to do the same task.

このようなことから、多義性解消の方策としては人間の認知的処理をモデル化したものではなく、計算機での処理に適したモデルを用いる場合が多い。例えば大量の電子化データが近年急速に入手し易くなったことを背景に、大量の例文集からの統計情報を利用した統計モデルや、大規模な機械可読辞書を利用した連想的な知識モデルなどを用いた多義性解消の試みが盛んに行われている[15]。

これらの試みのほとんどは、多義性解消に関する根本的な前提として、次の仮定に基づくものである。

仮定 2.1 同一の語義は、類似した文脈に現れる。

この仮定の下、単語の多義性は次の手順で解消することができる。

- あらかじめ多義語の各語義ごとに、語義選択のための手がかりとなる情報を与えておく。
- 新しく入力された多義語の文脈と与えられた情報との比較から、語義ごとに尤度を求める。

この二つの処理の結果、多義語の各語義には入力された文脈における尤度が割り当てられ、すなわち語義選択の優先順位が割り当てられることとなる。

2.3 先行研究

自然言語の解析では語義の曖昧性、すなわち単語の多義性が頻繁に出現する。一般に、辞書に定義されている単語のほとんどは多義語ではないが、実際に文に出現する単語はそのほとんどが多義語である。例えば、代表的な機械可読辞書であるWordNet[16]では、定義されている単語の80%以上が語義を一つしか持たない。しかし、WordNetに付随する例文集(corpus)であるWordNet Semantic Concordanceでは、出現する自立語(content word)のおよそ78%が多義語である。

多義性解消の問題は50年近くにわたって自然言語処理における最も基本的な問題の一つとして認識されており、これまでに多義性を解消するためのさまざまな手法が報告されている。これらの先行研究は、2.2.3節で言及した多義性解消の手順に従ったものがほとんどである。ここで、2.2.3節の手順に沿って多義性解消を行うときには、主に次の問題に対して妥当な解決を策定すればよいことになる。

- 語義選択の手がかりにどのような情報を用いるのか
- 語義選択の手がかりをどのように獲得するのか
- 尤度の算出方法など、語義選択をどのようにモデル化するのか

すなわち、多義性解消の手法はこれらの問題に対するアプローチの観点で特徴づけられる。

本節では、これまでに行われてきた先行研究のうち代表的なものについて概説し、各手法が上記の問題に対してどのようなアプローチを採択しているのかについて言及する。ここでは従来の手法を語義知識の観点と知識獲得の観点による二通りの分類を通して整理する。多くの場合、上記の三つの問題

に対するアプローチは互いに独立したものではなく、一つの問題に対するアプローチが他の二つに強く影響を及ぼす。従って以下、各手法が属する分類は、手法を最も強く特徴づけているアプローチによるものである。

2.3.1 語義知識による分類

語義知識による分類とは、語義選択の手がかりにどのような情報を用いるのかという問題に着目した手法の分類のことである。多義語の適切な語義を選択するには、多義語の文脈に存在する意味的な整合性を用いる。このとき、意味的整合性を判断する対象として、どのような情報を用いるのかを考えなければならない。この観点によって、従来の多義性解消の手法は連想関係に基づく手法(bag-of-word approach)と選択制限(selectional restriction)に基づく手法の二つに大別することができる。

2.3.1.1 連想関係に基づく手法

自然言語における単語の多義性の例として、次の文を考える[7]。

Treadmills attached to cranes were used to lift water from Roman times.

“crane”は「(重機としての)クレーン」の意味と「鶴」の意味を持つ多義語である。この多義性を人間が解決する場合、“crane”と文中の他の単語との意味的整合性から、重機としての意味を選択することができる。しかし、語義選択に強い影響力を持つと思われる単語“lift”は、“crane”に対して係り受け関係などの直接的な関係を持たない。“crane”と“lift”の間にある関係は、お互いがお互いを連想させる連想関係である。連想関係に基づく手法とは、このように多義語と連想関係を持つ単語を多義性解消の手がかりとする手法である。

連想関係に基づく手法は、多義語と連想関係を持つ単語を非順序集合として各語義に与え、これを語義選択の手がかりとする。新しく多義語が入力されたとき、多義語の周辺に現れる単語を抽出し、語義ごとに付与した連想関係を持つ単語が文脈にどの程度現れているかを求めて語義を選択する。すなわち、連想関係に基づく手法は多義語の文脈を多義語の周辺に現れる単語の非順序集合として扱うものである。通常、多義語の文脈はn-word windowの技法を用いて獲得される。

連想関係に基づく典型的な手法の一つとして、Yarowskyの報告[17]が挙げられる。Yarowskyの手法では、語義はロジェのシソーラス(Roget’s International

Thesaurus)から代表的なものとして選択された1042のシソーラスカテゴリによって定義される。コーパスにおいて、各カテゴリに属する単語の100-word windowから単語を抽出し、重み付き非順序集合にまとめたものが語義選択の手がかりとして見なされる。集合の各要素に付与される重みには相互情報量に類似したものが用いられる。多義語が入力されると、同様に多義語のn-word windowから単語が抽出される。抽出された各単語に対し、多義語が属するカテゴリごとに文脈情報との重複が調べられる。重複した単語に対してカテゴリごとに重みの総和が求められ、この値の最も高いカテゴリが語義として選択される。

語義知識として大規模コーパスにおける多義語のn-word windowを用いる手法には、Yarowskyの手法以外にもGaleらの手法[14]が良く知られている。Galeらの手法では、多義語の100-word windowから単語を抽出して多義語の各語義ごとに非順序集合にまとめ、抽出された単語に$\Pr(w_i|s_j)$のような条件付き確率を付与したものを語義選択の手がかりに用いる。ここで、$\Pr(w_i|s_j)$は語義s_jの100-word windowに単語w_iが出現する確率を意味している。新しく多義語が入力されたとき、語義s_jの尤度はあらかじめ獲得しておいた確率を用いて$\prod_{w \text{ in window of } s_j} \Pr(w|s_j)$と求められる。この尤度の最も高い語義が解として選択される。

コーパスなどの例文集から手がかりを獲得するのではなく、機械可読辞書から手がかりを獲得する試みも多く報告されている。例えば図2.5に示した辞書定義文の一部をよく観察すると、筆記用具としての“pen”の意味の定義文には“writing”,“drawing”,“ink”など、囲いの意味の定義文には“fence”,“farm animals”など、各語義と連想関係を持つ単語が含まれている。機械可読辞書を用いた手法の典型はLeskによる試みである[7]。Leskの手法は非順序集合として与えられた複数の多義語に対し、辞書定義文の間の重複が最大となる語義を選択するものである。例えば{“pen”,“sheep”}という多義語の集合が与えられたとき、“farm animals”という単語の重複によって語義“pen[1] 2”と語義“sheep 1”を選択することができる。

Véronis and Ideは辞書定義文から大規模なニューラルネットワークを構築するアプローチを報告している[19]。ネットワークにおける語義ノードはその定義文に含まれる単語ノードと活性リンクで結合され、同一の単語の語義ノード同士は抑制リンクで結合される。多義語を含む文が入力されると、入力文中の単語に対応する単語ノードを起点として活性値の伝搬がネットワーク上

pen[1] *n*

1 an instrument for writing or drawing with ink

2 a small piece of land enclosed by a fence, used for keeping farm animals in

3 **put/set pen to paper** to begin to write

4 *AmE slang* PENITENTIARY; a prison

pen[2] *v* **penned, penning** [T] *formal*

to write a letter or note with a pen

sheep *n* [C] *plural* **sheep**

1 a grass-eating farm animal that is kept for its wook and its meat

2 [often plural] someone who does not think independently, but follows what everyone else does or thinks

3 **separate the sheep from the goats** to find out which people are intelligent, skillful, successful etc, and which are not

4 **make sheep's eyes at** *old-fashioned* to look at someone in a way that shows you love them

図 2.5: 辞書の定義文の一部[18]

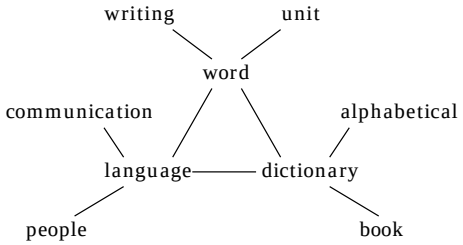


図 2.6: 参照ネットワークの一部[20]

に展開され、最終的に活性値の最も高い語義ノードが語義として選択される。Véronis and Ideの手法では語義選択の手がかりであるニューラルネットワークを構築する際に、辞書の見出し語と定義文に含まれる単語とをすべて同様に活性リンクで結合する。すなわち、辞書定義文に含まれる単語を非順序集合として扱っており、このことから手法は連想関係に基づく手法に分類される。

Niwa and Nittaは同様に辞書定義文からのネットワークを利用する手法を報告している[20]。このネットワークは参照ネットワークと呼ばれ、辞書の見出し語と定義文中の各単語とを図2.6のようにリンクで結合したものである。Niwa and Nittaの手法ではまず、頻度に基づいて選ばれた1000語の起点からのネットワーク上の距離に基づいて、各単語ごとに距離ベクトルを求めている。次にコーパスにおける各語義のn-word windowに含まれる単語に対して距離ベクトルを合成し、これを各語義の文脈情報として利用する。入力された多義語に対しても同様にベクトルが合成され、ベクトルの内積が最も高い語義を選択する。

2.3.1.2 選択制限に基づく手法

語義選択のための意味的整合性判断の対象として、最も強い影響力を持つと考えられるのが、多義語と直接的な構文関係を持つ単語である。例えば、次の二つの文を考える。

Susan opened the meeting.
Susan opened the door.

この例の “open” はそれぞれ、「(会議などを)開催する」という意味と、「(窓・戸などを)開ける」という意味で用いられている。それぞれの文脈に従って “open” の語義を選択する際に連想関係に基づく手法は、「“open” のn-word windowに

“meeting” が含まれる」あるいは「“open” の n-word window に “door” が含まれる」といった情報を用いる。しかし、連想関係に基づく情報では

Susan opened the door of meeting room.

という例に対して、“door” と “meeting” を同じ重みで扱ってしまう。“open” の語義選択に最も強く影響を与えるのは、“What did Susan open?” の観点、すなわち “open” の直接目的語は何かという観点からの情報である。

一般に、単語の格充足性に関する意味的制約を選択制限と呼ぶ。例えば、「(会議などを) 開催する」という意味の “open” は、直接目的語に会議などの事象を置くという制限を持ち、「(窓・戸などを) 開ける」という意味の “open” は、直接目的語に入り口の役割を持つ物体を置くという制限を持つ。選択制限に基づく手法とは、こうした制限に基づいて多義語と直接的な構文関係を持つ単語を多義性解消の手がかりとする手法である。

選択制限に基づく手法は、多義語と特定の構文関係を持つ単語とその関係の種類を各語義に与え、語義選択の手がかりとする。新しく多義語が入力されたとき、与えられた構文関係を持つ単語を多義語の文脈から獲得し、得られた単語が選択制限を満たすものかどうかを調べることによって語義を選択する。すなわち、文脈を多義語と特定の構文関係を持つ単語の観点で扱うという特徴を持つ。用いられる構文関係としては、動詞と目的語、名詞と修飾語などの関係が考えられる。

Brown らは多義語の持つ意味をあらかじめ高々二つに限定し、ある特定の関係を持って多義語の周辺に現れた単語の一つに対して binary question を設問することによって多義性解消を試みている [21]。binary question とは、「多義語と特定の関係を持つ単語は w という単語か」というような質問であり、“binary” はこの回答が是か否かの二値であることに由来している。binary question の対象となり得るのは

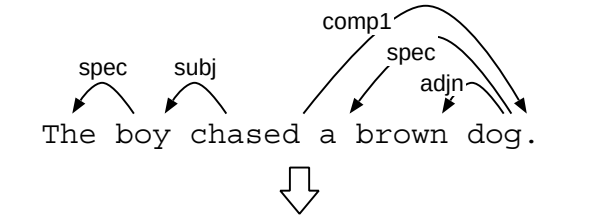
- 多義語自身
- 多義語の左隣の単語
- 多義語の右隣の単語
- 多義語の左側に走査していったときに最初に現れる名詞
- 多義語の右側に走査していったときに最初に現れる名詞

- 多義語の左側に走査していったときに最初に現れる動詞
- 多義語の右側に走査していったときに最初に現れる動詞

の七つであり、このうち多義語に最も有意な手がかりを与えるものを選択して設問する。手がかりの選択はエントロピーの尺度に基づいており、選択された手がかりに従ってコーパスから語義ごとに単語が抽出される。入力された多義語の文脈から同様に抽出された単語に対し、あらかじめ与えられた集合に含まれるかどうかを検査することによって多義性解消が行われる。

Hearstはpartial parsingの結果を利用して多義語と構文関係を持つ単語を頻度と共にコーパスから抽出し、これを手がかりに名詞の多義性解消を行う手法を報告している[22]。Hearstの手法で語義選択の手がかりとして抽出される単語とは、多義語を修飾する単語、多義語が修飾する単語、多義語に隣接する句の主辞などである。また、これ以外に多義語の語彙的な情報も手がかりとされており、多義語の頭文字が大文字にされている回数や、多義語が他の単語を修飾する回数などが手がかりとして抽出される。多義語が入力されると、入力文からも同様に手がかりが抽出され、あらかじめ与えられた手がかりとの重複が調べられる。重複した手がかりに割り当てられた頻度が語義ごとに合計され、合計値の最も高いものが語義として選択される。

Linは多義語と構文関係を持つ単語を関係の種類と共にコーパスから抽出し、これを局所的文脈と見なすことで多義性解消を試みている[23]。Linの手法は多くの手法で用いられている仮定2. 1とは僅かに異なる仮定を用いて多義性解消を試みている。その仮定とは、「同一の局所的文脈を持つ単語同士は類似した意味を持つ」というものである。Linは任意の単語wに対して、構文関係の種類、wと構文関係を持つ単語、主辞・修飾辞の別、の三つ組を定義し、これを局所的文脈(local context)と呼んでいる。図2.7にLinの定義した局所的文脈の例を示す。ここで、各三つ組における`subj`、`adjn`、`comp1`は構文関係の種類を表しており、それぞれ主語(subject)、修飾語(adjunct)、第一補語(first complement)を意味する。Linの手法では、同一の局所的文脈ごとに対応する単語を収集し、これを語義選択の手がかりとする。新しく多義語が入力されると、多義語の局所的文脈と同一の局所的文脈に対してあらかじめ獲得しておいた単語の集合を求める。この集合から語義同士の類似度に基づいたスコアが求められ、最も高いスコアを持つ語義が選択される。



Word	Local Contexts
boy	(subj chase head)
dog	(adjn brown mod) (comp1 chase head)

図 2.7: 局所的文脈の例 [23]

2.3.2 学習方法による分類

学習方法による分類とは、語義選択の手がかりとなる情報をどのように学習するのかという知識獲得の問題に着目した手法の分類のことである。2.2.3 節の手順に従った多義性解消では、語義知識を多義語の各語義ごとに与える必要がある。語義知識の獲得源としてコーパスを考えたとき、知識を獲得しようとしているコーパス自身が曖昧であるという問題に突き当たる。この問題は知識獲得障害 (knowledge acquisition bottleneck) の問題としてしばしば言及される。

この問題を回避する直観的な方法は、訓練データ中の多義語に人手であらかじめ語義情報を付与しておくというものである。訓練データにおける多義語の語義を明らかにした状態で語義知識を学習する手法は教師付き学習 (supervised learning) と呼ばれる。訓練データとしてあらかじめ語義情報が付与されている機械可読辞書を獲得源に用いる学習も教師付き学習に分類される。一方、多義語の語義が明らかでない状態での学習は教師なし学習 (unsupervised learning) と呼ばれる。多義性解消の手法は教師付き学習か教師なし学習かの観点で分類することができる。

2.3.2.1 教師付き学習に基づく手法

教師付き学習に基づいてコーパスから語義知識を獲得するには、訓練コーパス (training corpus) における多義語のすべての出現にあらかじめ語義情報を人手で付与しておかなければならない。人手による形態素情報や構文情報の付与に比べ、語義情報の付与に関しては作業効率向上のための環境構築が未

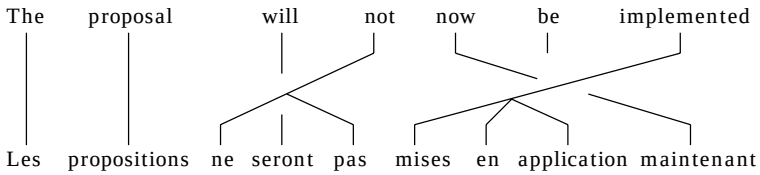


図 2.8: 単語の対応付けの例 [24]

だ不充分なこともあり、語義情報付与にかかるコストは極めて大きい。このため、従来語義情報が付与されたコーパスは規模が小さいものか、語義情報を付与する多義語を一部の特定のものに限定したものが多く、手法の大規模な評価は困難な傾向にあった。

こうした問題への取り組みとしてよく知られているのが、2.3.1.2節で概説したBrownらの手法で採用されている試みである。Brownらは対訳テキストを用い、一方の言語の多義性を他方の語の情報を利用することで解消し、これによって語義情報を付与した訓練データを自動生成する試みを報告している [24]。例えば、英単語 “duty” は「税」の意味と「義務」の意味を持つ多義語である。これと同じ多義性を持つ単一の仏単語は存在せず、ほとんどの場合 “duty” は “droit” (税) か “devoir” (義務) のどちらかに仏訳される。従って、仏訳が “droit” である “duty” の語義は「税」の意味であり、“devoir” であるものは「義務」の意味である。Brownらはこの点に着目し、英仏の並列コーパス (parallel corpus) であるカナダ議会の議事録 Hansards を利用している。この観点で語義情報付きの訓練データを作成する際には、図2.8のように特定の英単語の仏訳がどの仏単語に対応するか (あるいはその逆) を明らかにする必要がある。Brownらは一文中の単語数や anchor point と呼ばれる情報を用いて単語の対応付けを行う手法を報告しており、実験を通じておよそ99%の高い精度で単語の対応付けが可能であることを示している。Brownらの多義性解消の手法は、こうした手続きを経て得られた訓練データから手がかりを獲得しており、従って大規模な教師付き学習を実現している。

一方、大規模な電子化コーパスの重要性が近年さまざまな方面で認知されており [25]、これに伴なって実用的な規模の語義情報付きコーパスがいくつか入手可能になっている (WordNet Semantic Concordance、EDR Corpus[26] など)。語義情報付きコーパスを用いた手法の一つに、Stetina and Nagao の手法 [27] がある。Stetina and Nagaoの手法では、文中の構文関係は六つ組 ($PNT, MNT, HNT,$

MS, HS, RP) によって定義される。ここで、構文構造の生成規則はチョムスキーの標準形 (Chomsky's normal form) で記述される ($A \rightarrow BC|a$ $(B, C \in N, a \in \Sigma)$) ものとし、六つ組の各記号はそれぞれ規則の左辺の非終端記号、右辺の修飾語に相当する非終端記号、右辺の非修飾語に相当する非終端記号、修飾語の語義、非修飾語の語義、修飾語と非修飾語の相対的位置関係を表す。このとき、二つの単語が特定の語義の組み合わせで特定の構文関係を持つ確率を利用することによって多義性を解消する。構文関係を表す六つ組には、構文関係を持つ二つの単語の語義が含まれており、この確率は語義情報が付与されたコーパスから教師付き学習で獲得する必要がある。

2.3.2.2 教師なし学習に基づく手法

一般に教師付き学習に基づく手法は、語義情報付き訓練データのコストの観点から学習の規模や手法の評価の規模が比較的小さなものになる傾向がある。教師なし学習は、教師付き学習に内在するこうした問題を解消するものであり、訓練データのコストを単に電子化されただけのテキストデータ (raw corpus) のコストと同一の水準にまで下げることができる。大規模な訓練データを容易に入手できることから、教師なし学習に基づく多義性解消に関する研究は盛んである。但し、訓練データからの学習が完全に正当に行われるとは言えないため、教師なし学習は一般に教師付き学習に比べて学習の精度が低く、獲得できる語義情報の品質は教師付き学習に劣る。

本論文で定義する語義の割り当ての観点では、完全な教師なし学習に基づく多義性解消は厳密に言えば本質的に不可能である。なぜならば、訓練データにおける出現と語義とを対応づける情報が存在しないためである。また、完全な教師なし学習に基づく手法は一般に一切の外部情報を用いないことを前提とするものが多く、従って各単語がどのような語義を持つのかという情報も利用できない。このため、教師なし学習に基づく手法は語義の割り当てではなく、語義の弁別の観点で多義性解消を行う。すなわち、教師なし学習に基づく多義性解消の問題は、多義語の任意の二つの出現に対し、両者が同じ語義を持つかどうかを両者の文脈から判断するクラスタリング (clustering) の問題に帰着される。

こうした観点に基づく典型的な手法としてSchützeの手法 [28] を挙げることができる。Schützeの手法では、単語の多義性解消の問題は単語の出現回数を各次元の要素としたベクトル空間におけるクラスタリングの問題に帰着され

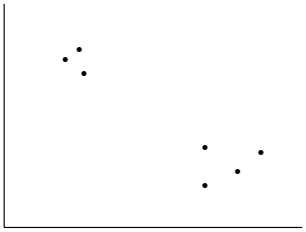


図 2.9: 文脈ベクトルのクラスタ

る。ベクトルで表現されるのは単語と文脈と語義である。単語ベクトルとは、特定の単語のn-word windowに他の各単語が出現する回数を次元の要素としたものである。文脈ベクトルは特定のn-word windowに出現する単語の単語ベクトルを合成したものである。語義ベクトルはSchützeの手法において語義の表現として用いられるものであり、図2.9のような文脈ベクトルのクラスタに含まれるベクトルを合成したものである。文脈ベクトルのクラスタリングにはEMアルゴリズムと呼ばれるクラスタリングアルゴリズムが用いられる。すなわち、Schützeの手法では語義知識として、あらかじめ訓練データから教師なし学習でクラスタを求め、そのクラスタから語義ベクトルを合成したものを用いる。新しく多義語が入力されたとき、多義語の文脈ベクトルと語義ベクトルを比較し、最も類似した語義ベクトルを多義語の語義とする。類似性判断にはベクトルの内積が用いられる。

福本らは動詞をn次元の名詞空間におけるベクトルで表現し、ベクトルのクラスタリングに基づく動詞の多義性解消を報告している[29]。動詞vのベクトルの要素はvとn個の名詞それぞれとの相互情報量である。福本らの手法は、類似した語義を持つ二つの動詞が同一の名詞と共起(co-occurrence)して現れるという仮定に基づいて、多義語の語義を別の動詞の語義に同定している。この観点の下、訓練データにおいて多義語のn-word windowに現れる名詞を用いて多義語の出現ごとにベクトルを生成する。このとき、得られたベクトルは多義語の持つ複数の語義情報が合成されたものと見なし、図2.10のようにベクトルを各軸に従って分割し、分割された各ベクトルに対して他の動詞(クラスタ)とのクラスタリングを試みる。クラスタリングには集合の偏差に基づく値が用いられる。新しくクラスタが生成された場合、多義語はその動詞と同じ語義を持つと見なすことができる。また、このときの軸に相当する名詞は手がかりとして獲得され、新しい多義語の文脈から得られた名詞と重複が調

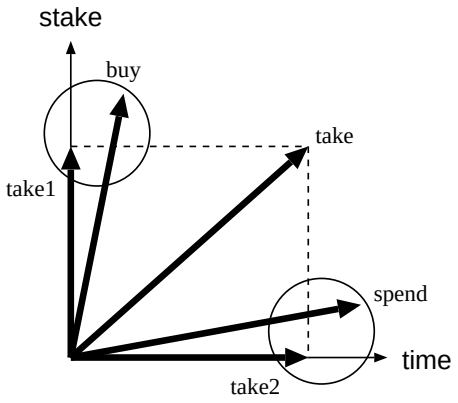


図 2.10: ベクトル “take” の分割 [29]

べられる。福本らの手法ではこのときの重複に従って語義が選択される。

また、小数の訓練データだけにあらかじめ人手で語義情報を付与し、その情報を利用して徐々に利用できる訓練データの量を増やすブートストラップ (bootstrap) のような手法で教師なし学習を実現している手法もある。この手法の代表的なものとして Yarowsky の手法 [30] を挙げることができる。Yarowsky の手法では、まず訓練データにおいて多義語の各語義が出現する典型的な連語表現 (collocation) に人手で語義を割り当てる。この連語表現から “One Sense Per Collocation” の性質を用いて、訓練データに現れる同様の連語表現に同一の語義を割り当てることができる。この段階で語義情報が付与された多義語は訓練データ中のすべての多義語の出現の 1%程度を占める。次に、語義情報が付与されたデータを用いて決定リストの教師付き学習を行い、この決定リストを用いて全訓練データに対する多義性解消を試みる。このときの多義性解消は多義語の語義に尤度を割り当てるものであり、解として選択された語義の尤度が特定の閾値 (threshold) を超えたものだけを信頼できるものとして採択し、訓練データに割り当てる。超えない例に対しては語義の割り当てを行わない。得られた訓練データに対し、“One Sense Per Discourse” の性質 [31] を用いてさらに語義情報を割り当てることができる。これらの処理を繰り返すことによって訓練データのほぼすべてに語義情報を割り当てることができる。

2.4 本研究の位置付け

語義知識の観点から多義性解消を考えたとき、多義性解消の精度向上を妨げている要因として連想関係と選択制限の親和性の低さを指摘することができる。一般に、連想関係に基づく手法は単語間の連想関係を言わば選好の条件として利用しているのに対し、選択制限に基づく手法は構文関係を持つ単語間の意味的整合性を言わば制約として利用している。語義選択の手がかりとしては対照的な観点に基づくものであるため、両者の手がかりを組み合わせて多義性解消を試みている研究は極めて少ない。

連想関係、選択制限それぞれに基づく手がかりを組み合わせて用いる手法に分類されるものとして、例えばYarowskyの手法[32]が挙げられる。Yarowskyの手法では、連想関係と選択制限のそれぞれに基づく手がかりを独立に階層的決定リストのノードに割り当てている。それぞれの手がかりには、多義語の各語義と手がかりとの共起頻度に基づいた重みが割り当てられ、これに基づいて決定リストの順序が求められる。多義語の語義は、得られた決定リストを辿ることで選択される。Yarowskyの手法は連想関係、選択制限それぞれによる手がかりを決定リストのノードにおいて独立に利用しており、各手がかりは段階的に利用されるため、厳密に両者を組み合わせた情報を用いているわけではない。

本研究は従来より精度の高い多義性解消の実現を目的とするものである。多義性解消の精度向上に対する本研究の基本的な考え方は、語義選択の手がかりとして連想関係と選択制限を組み合わせた情報を用いることである。この考え方に基づき、本研究では連想関係に基づく情報と選択制限に基づく情報とをそれぞれ独立に扱った場合に、多義性解消が困難になる事例を明らかにする。こうした事例を適切に処理するため、本研究では連想関係と選択制限を両者ともよく保存する文脈の表現形式を構築し、これに沿って訓練データから収集された情報を語義選択の手がかりとする。また、この情報を用いて語義の尤度を妥当な値に数値化するための手法を構築する。なお、本研究では問題を単純にするために対象とする多義語の品詞を動詞に限定した。

一方、こうした語義選択の情報を訓練データから獲得することを考える。一般に、機械可読辞書から獲得できる情報は辞書項目と定義文の間の連想関係であり、選択制限を獲得することは困難と言える。本研究では語義知識の獲得源に大規模コーパスを用いる。2.3.2.1節で言及したように、多義性解消の手

がかりとなる情報の獲得源として実用的規模の語義情報付きコーパスを求めることが近年急速に容易になりつつある。この背景を鑑みると、教師なし学習を行うことによる語義知識の品質低下は回避すべき問題である。

学習方法による多義性解消の精度の違いを認識できるものとして、多義性解消の国際的なコンテストであるSENSEVAL[33]がある。SENSEVALは多義性解消のさまざまな手法に基づく多数のシステムが参加して1998年に開催された。SENSEVALに参加したシステムには、教師付き学習に基づく多義性解消の手法を用いたシステムと教師なし学習に基づくシステムとの両者ともが多数含まれる。SENSEVALではこれらのシステムを同一の評価基準で評価しており、すなわち、さまざまな手法を一元的に評価した点で極めて興味深い結果を残している。SENSEVALの開催報告[34]によると、語義情報付きの訓練データが利用可能であった場合、それを利用したシステム、すなわち教師付き学習に基づくシステムは一般に教師なし学習に基づくものよりも精度の面で大きく上回る傾向がある。こうした点からも教師なし学習と比較して教師付き学習が優位にあることが認識できる。

以上の理由から、本研究では語義知識の獲得に語義情報付きコーパスからの教師付き学習を採用する。語義知識の獲得形態として教師付き学習を採用することは、多義性解消の精度向上に大きく寄与すると期待できる。

本研究の位置付けは本章を通して次のようにまとめることができる。

本研究の大域的位置付け 本研究は意味解析の一部として位置付けられる単語の多義性解消について、従来よりも高い精度で多義性を解消する手法を構築するものである。

本研究の局所的位置付け 本研究では、語義選択の手がかりとしてあらかじめ多義語の各語義ごとに与えられる情報に、連想関係と選択制限を組み合わせた情報を用いる。こうした情報を用いた多義性解消に関する研究は従来ほとんどなく、これによって従来の手法よりも高い精度の多義性解消を実現する。

学習の観点からの位置付け 実用的規模の語義情報付きコーパスが比較的容易に入手できるという背景から、本研究では語義選択の手がかりとなる情報を、語義情報付きコーパスからの教師付き学習で獲得する。これによって精度の高い語義知識を獲得することが期待できる。

第 3 章

ペアワイズアライメントを用いた動詞の多義性解消

本章は、本研究で構築した動詞の多義性を解消するための新しい手法について説明するものである。単語の多義性の問題は古くから自然言語処理における最も重要な問題の一つとして位置付けられており、これまでに様々な多義性解消の試みが報告されている。従来の試みは多義語の文脈の扱いの観点から、多義語の周辺の単語を非順序集合として用いるもの(連想関係に基づく手法)と、構文関係を用いるもの(選択制限に基づく手法)の二つに大別できる。しかし、これらの手法はそれぞれ異った観点で手がかりを求めており、精度の向上には限界が考えられる。本章で提案する手法は、多義語の文脈として一文の依存構造木全体を用いており、二つの手法の特長を併せ持つものである。本手法では、DNA配列の類似性評価に広く用いられているペアワイズアライメントの技法に基づいて文脈の類似性を評価する。これによって、文脈間の類似度を柔軟かつ頑健に求めることが可能である。本手法は人手による教師付き学習を必要とするが、多義性解消の実験からは平均81.1%の精度が得られた。

3.1 はじめに

多義性解消とは多義語の適切な意味を文脈から推定することである。その応用範囲は多岐に亘り、機械翻訳における訳語の選択や情報検索における検索結果の絞り込みなどが挙げられる。単語の多義性の問題は50年近くに亘り自然言語処理における最も基本的な問題の一つとして認識されている。

2.3.1節で述べたように、これまでに提案されている多義性解消の手法のほ

とんどは文脈の扱いによって大きく二つに分類できる。

連想関係に基づく手法 文脈を多義語の周辺に現れる単語の非順序集合として扱う。n-word windowに基づいた手法に代表される。

選択制限に基づく手法 文脈を多義語と特定の構文関係を持つ単語の観点で扱う。例えば動詞と目的語、名詞と修飾語などの構文関係が用いられる。

2.4節で言及したように、これらの手法の文脈の扱いは対照的な観点に基づく手がかりを与えるものである。連想関係に基づく手法は多義語の文脈を単語の非順序集合で表現するため、単語間の構文関係が考慮できないという問題がある。一方、選択制限に基づく手法は多義語と直接的な構文関係を持つ単語を手がかりとして用いることはできるが、直接的な構文関係を持たない単語を考慮できない。すなわち、一文中の一部の単語だけしか考慮できないことが問題である。従って、連想関係に基づく手法は選択制限を考慮することができず、選択制限に基づく手法は連想関係を考慮することができないという性質が存在し、これらの手法では精度の向上に限界がある。

例えば次の文において動詞 “fire” の語義を “go off or discharge” と “terminate the employment” のどちらかに決定する場合を考える[1]。ここでは簡単のため一文のみの文脈を考えるものとする。

> My Cousin Simmons carried a musket, but he had loaded it with bird shot, and as the officer came opposite him, he rose up behind the wall and fired.

“musket”、“loaded”、“bird shot” などの単語は “fire” の語義を “go off or discharge” に導く有用な手がかりと考えられる。“terminate the employment” に導く単語は特に見当たらない。従ってこの例文は連想関係に基づく手法で適切に処理できる。しかし、手がかりと思われる単語は “fire” と直接的な構文関係を持たないため、選択制限に基づく手法では語義選択が困難である。

一方、次の例文を考える。

> Police said Haga was immediately fired from the force.

この文では “Haga”(人名)が “fire” の目的語として現れている点が重要な手がかりと考えられる。選択制限に基づく手法はこの直接的な構文関係を利用して

[1] ここで示す二つの例文はそれぞれBrown Corpus (WordNet Semantic Concordance)、EDR Corpusより引用した。

適切な語義選択が可能である。しかし、連想関係に基づく手法では不適切な意味(“go off or discharge”)を導く単語(“Police”や“force”など)を排除できず、手がかりを適切に用いることが困難である。これは文脈を単語の非順序集合として一様に扱ってしまうためである。

こうした不都合を解決し、多義性解消の精度をより向上させることを目的として、本研究では教師付き学習に基づく多義性解消の新しい手法を構築した。本章は本研究で構築した手法について詳述するものである。本手法は手がかりとして一文全体の依存構造木から抽出される単語の配列を用いる。3.3節で述べるように、単語の配列は一文中のすべての単語を含み、また一文中の係り受け関係を反映した構造を持つ。従って連想関係、選択制限を双方とも考慮した手がかりが提供できる。また、単語の配列の導入に伴って、本手法はDNA配列や蛋白質のアミノ酸配列の類似度推定に広く用いられているペアワイズアライメントの技法を用いている。本手法については3.3節で基本的な考え方を述べた後、3.4節で詳述する。3.5節では本手法を用いてSENSEVALの動詞を対象に行った多義性解消に関する実験について述べる。

3.2 準備

3.2.1 単語の配列

本小節では本研究で構築した手法で用いる単語の配列と、それに付随したいくつかの用語について定義する。

定義 3.1 任意の文Sにおける単語の順序対(w, w')において、wを被修飾語、w'を修飾語とする。このとき、wとw'の間の関係を係り受け関係(依存関係; dependency)といい、wを主辞(head)、w'を修飾辞(modifier)という。∎

定義 3.2 任意の文Sに含まれるすべての単語の集合をVとし、すべての係り受け関係の集合をEとする。ここで、Eの要素は係り受け関係を持つ単語の順序対であり、主辞から修飾辞への方向を持つものとする。このとき、二つ組(V, E)を依存構造木(dependency tree)という。∎

定義 3.3 依存構造木(V, E)における任意の単語$w \in V$に対し、集合$\{(w, w') | w' \in V, (w, w') \in E\}$の個数を$O_w$で表し、$\{(w', w) | w' \in V, (w', w) \in E\}$の個数を$I_w$で表すとする。このとき、$I_w = 0$を満たす$w$は唯一存在し、これを根(root)という。また、$O_w = 0$を満たす$w$を葉(leaf)という。∎

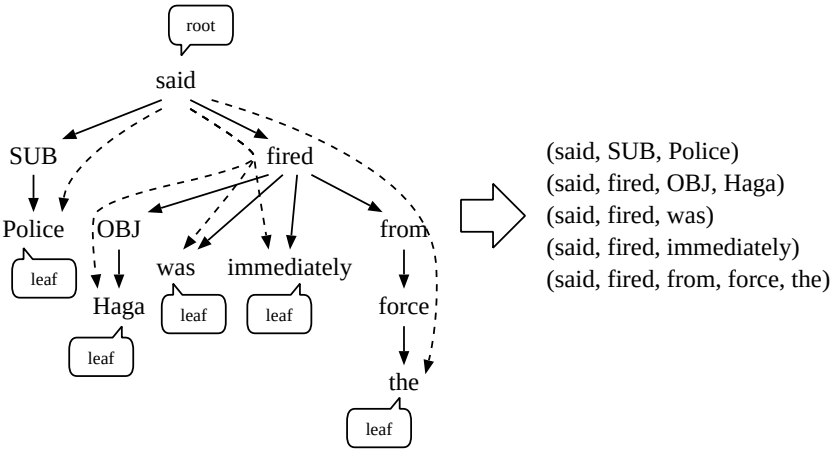


図 3.1: 依存構造

定義 3. 4 依存構造木(V, E)における単語と係り受け関係の交互列によって構成される順序集合 $(w_0, (w_0, w_1), w_1, \cdots, (w_{m-1}, w_m), w_m)$を$w_0$から$w_m$への経路(path)という。 ∎

定義 3. 5 依存構造木(V, E)における根から葉への経路を単語の配列(word sequence)という。 ∎

任意の依存構造木(V, E)は$I_W = 0$を満たす唯一の根$W \in V$を持ち、W以外の任意の単語$w \in V$は常に$I_w = 1$を満たす。また、任意の経路$(w_0, w_1 \cdots, w_m)$に対して常に$w_0 \neq w_m$である。従って(V, E)は有向木(directed tree)であり、任意の経路に含まれる係り受け関係は経路に含まれる単語によって一意に決定できる。このことから、単語の配列は係り受け関係を省略した順序集合$(w_0, w_1, \cdots, w_m)$で表すことができる。

図3.1に依存構造木の例を示す[2]。図3.1の依存構造木では根は“said”であり、葉は“Police”、“Haga”、“was”、“immediately”、“the”の五つである。従ってこの例では図3.1の右側に示す単語の配列が得られる。

3.2.2 ペアワイズアライメント

アライメント(alignment)とは、任意のk個の配列において配列要素間の最適な対応付けを求める技法である[35, 36]。$k = 2$のアライメントを特にペア

[2]図3.1では依存構造木にノード“SUB”、“OBJ”が追加されている。これは動詞の主格と目的格の違いを明確にするために追加するノードであり、“SUB”は主格、“OBJ”は目的格を表している(付録参照)。

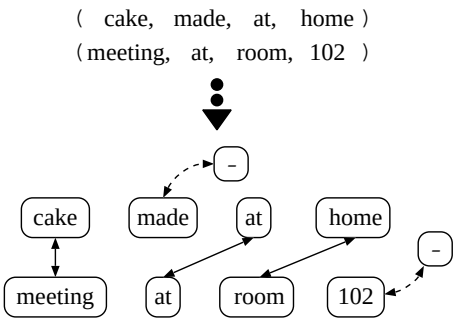


図 3.2: ペアワイズアライメント

ワイズアライメント (pairwise alignment) という。配列要素間の対応付けを求める際には、対応関係の非交差を条件とする。従って、対応関係が求められない配列要素が存在する (図3.2)。このとき、対応関係のない配列要素については便宜的にギャップ (gap) と呼ばれる記号 "–" と対応しているものと見なす。図3.2の場合、"made", "102" は適切な対応関係が求められず、ギャップと対応付けられる。ギャップの概念を用いることにより、ペアワイズアライメントは二つの単語の配列 $p = (w_{p,1}, w_{p,2}, \cdots, w_{p,m})$、$q = (w_{q,1}, w_{q,2}, \cdots, w_{q,n})$ を同一の長さ $N(\geq m, n)$ を持つ配列対 $p' = (w_{p',1}, w_{p',2}, \cdots, w_{p',N})$、$q' = (w_{q',1}, w_{q',2}, \cdots, w_{q',N})$ に変形する操作と考えることができる。p'、q' はそれぞれ p、q に適宜ギャップを挿入することで求められる配列である。配列要素の対応付けは変形配列 p'、q' の同一の位置にある要素間で行われるものと見なす。すなわち、対応付けられる配列要素対は $(w_{p',i}, w_{q',i})(1 \leq i \leq N)$ である。

以下ではペアワイズアライメントの定式化に必要ないくつかの概念について定義する。

定義 3. 6 任意の単語対 (w, w') が WordNet においてそれぞれ概念ノード (synset) $s_1, s_2, \cdots, s_m$、$s'_1, s'_2, \cdots, s'_n$ を持つとする。このとき、単語対 (w, w') の対応付けの評価値 $d(w, w')$ は次式で与えられる。

$$d(w, w') = \max_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} (1 - 2 \cdot (sd(s_i, s'_j))^2)$$

∎

この式は単語対 (w, w') 間の類似度を表すものであり，$sd(s_i, s'_j)$ は概念ノード s_i と s'_j の間の Semantic Distance[27] を表す[3]。Semantic Distance は、WordNet に

[3]本定義で用いる Semantic Distance は名詞、形容詞、副詞に関しては Stetina and Nagao の

おける共通の上位ノードまでの距離、類義関係、対義関係などに基づいて二つの概念ノード間の距離を推定する尺度である。定義3. 6では、概念ノード間の距離の差をより顕著にするため、Semantic Distanceを2乗している。$0 \leq (sd(s_i, s'_j))^2 \leq 1$であるため、$-1 \leq d(w, w') \leq 1$である。

定義 3. 7 任意の単語wとギャップの対応付けの評価値をギャップスコアという。ギャップスコアは次式で定義される。

$$d(w, \text{“-”}) = d(\text{“-”}, w) = -1$$

∎

定義 3. 8 任意の配列p, q間のアライメントスコア$AS(p, q)$を次式で定義する。

$$AS(p, q) = \max_{p', q'} \sum_{i=1}^{N} d(w_{p',i}, w_{q',i})$$

但し，任意の配列aの要素数を$|a|$で表すとき，$N = |p'| = |q'|$である. ∎

これらの定義を用いて、ペアワイズアライメントは次のように定式化できる。

$$(p', q') = \arg\max_{p', q'} \sum_{i=1}^{N} d(w_{p',i}, w_{q',i})$$

最適なペアワイズアライメントを求めるアルゴリズムとしては、動的計画法に基づく手法、有限状態オートマトンに基づく手法、隠れマルコフモデルに基づく手法などが良く知られている[38]。本手法では、単語の配列の長さが一様でないことを考慮し、得られたペアワイズアライメントの左右両端に位置するギャップにペナルティを与えないアルゴリズムを用いている(付録参照)。

3.3 基本的な考え方

本研究の多義性解消に対するアプローチは、多くの先行研究と同様に2.2.3節に言及した仮定2. 1 に基づいている。2.2.3節で述べたように、多義性解消の問題はこの仮定の下、多義語の文脈表現の問題と文脈の類似性評価の問題の二つに分割して考えることができる。本研究の多義性解消に対する基本的な

定義に従うが、動詞に関しては概念ノード単位のマッチングのみを行う。すなわち、動詞についての階層構造を参照しない。WordNetでは、名詞の階層構造が上位下位関係を基に構築されているのに対し、動詞の階層構造は様態(manner)の継承を基に構築されている[37]。このことから、WordNetの階層構造を用いた動詞概念間の直観的な距離推定は困難である。

考え方は、前者に対して単語の配列を用い、後者に対してペアワイズアライメントの技法を用いるものである。

任意の文をSとし、Sの依存構造木を(V, E)とする。3.1節で述べたように、従来の多義語の文脈表現は多義性解消の観点からは不満である。すなわち、連想関係に基づいた文脈表現では文中の係り受け関係が考慮できず、Eに保存される手がかりが欠失する。また、選択制限に基づいた文脈表現では文中の一部の係り受け関係しか考慮できず、Vに保存される手がかりが欠失する。3.1節の二つの文は、この問題のためにそれぞれの手法が適切に多義性解消できない例を示すものである。これに対し本研究は、文脈を単語の配列を用いて表現する。依存構造木(V, E)から得られる配列の集合を$P_S = \{p_{S,1}, p_{S,2}, \cdots, p_{S,N}\}$とするとき、単語の配列に関して次のことが成り立つ。ここで、$p_{S,i}$は単語の配列を表す。

- $p_{S,i}$に含まれる単語の集合を$V_{p_{S,i}}$で表すとき、$\bigcup_{1 \leq k \leq N} V_{p_{S,k}} = V$である。
- $p_{S,i} = (w_1, w_2, \cdots, w_n)$の要素から隣り合う任意の単語対$(w_k, w_{k+1})$を抜き出したとき、$1 \leq k \leq n-1$の任意の$k$に対して$(w_k, w_{k+1}) \in E$である。
- $p_{S,i}$に含まれるすべての隣り合う単語対を$E_{p_{S,i}}$で表すとすると、$\bigcup_{1 \leq k \leq N} E_{p_{S,k}} = E$である。

すなわち、単語の配列は文中のすべての単語Vとすべての係り受け関係Eを両者ともよく保存する。

文脈表現に単語の配列を用いるとき、文脈の類似性評価の問題に対して集合要素のマッチングに基づく方法を適用することは現実的でない。なぜならば配列は複数の単語の順序集合であり、同じ配列が現れる確率が極めて低いためである。本研究では、ペアワイズアライメントによって求められる配列間の類似度に基づいて文脈の類似性を評価する。単純なマッチングに基づいた評価では配列一致の可否を判断基準とするのに対し、ペアワイズアライメントに基づいた評価では配列の類似の度合いを判断するため、頑健な評価が可能である。ペアワイズアライメントは次の性質を持つ。

性質1 アライメントスコアは対応付けられた単語間の類似度の総和であるため、配列に含まれる単語$V_{p_{S,i}}$の類似性を反映する。

性質2 対応関係の非交差の制約から、アライメントスコアは係り受け関係$E_{p_{S,i}}$の類似性を反映する。

性質3 アライメントは配列変形の最適解を求めることと等価であるため、評価値(アライメントスコア)を頑健に求めることができる。

これらの性質は、単語の配列の類似性評価にペアワイズアライメントを適用することの妥当性を示唆している。

3.4 提案する手法

本研究で構築した多義性解消のアルゴリズムを以下に示す。

多義性解消のアルゴリズム

Step 1 多義語wの語義$s_1, s_2, \cdots, s_n$に対し、訓練データから語義ごとに配列パターンの集合$P_1, P_2, \cdots, P_n$を生成する。

Step 2 多義語wを含む入力文S_wを構文解析し、依存構造木(V, E)を獲得する。

Step 3 (V, E)から単語の配列の集合Q_{S_w}を求める。

Step 4 P_iとQ_{S_w}との類似度を$sim(P_i, Q_{S_w})$とするとき、
$k = \arg\max_i sim(P_i, Q_{S_w})$を満足する語義$s_k$を解として選択する。

Step 1で生成する配列パターンとは、訓練データから語義ごとに収集された単語の配列をパターン化したものである。配列パターンについては3.4.1節に詳述する。Step 2は既存の構文解析システムを用いて入力文を構文解析することを意味しており、本手法では構文解析システムから正しい解析結果が得られることを仮定している。Step 3はStep 2の解析結果から単語の配列を定義3. 5 に基づいて抽出することを意味している。Step 4では、語義ごとに獲得した各パターン集合$P_1, P_2, \cdots, P_n$と入力文から得られた配列の集合Q_{S_w}との類似性を評価する。類似性評価はペアワイズアライメントに基づいており、評価方法は3.4.2節に詳述する。

3.4.1 配列パターン

本節ではStep 1で生成する配列パターンについて述べる。配列パターンとは、訓練データから語義ごとに収集された単語の配列をパターン化したものである。生成したパターンは語義ごとに集合として与えられ、得られたパターン集合が各語義の文脈情報となる。

Q_{S_1} : (appended,OBJ,close-up,of,one,of,band,firing,straight)
(appended,OBJ,close-up,of,one,of,band,*firing,OBJ,revolver*,his)
(appended,OBJ,close-up,of,one,of,band,firing,at,audience,the)
...

Q_{S_2} : (fired,into,bed,the)
(*fired,OBJ,slugs*,several)
(fired,were)
...

Q_{S_3} : (was,fired,been,have)
(was,*fired,OBJ,shot*,the)
(was,fired,from,hotel,either)
...

Q_{S_4} : (fired,are)
(*fired,OBJ,guns*,the)
(fired,carry,to,vault,the)
...

図 3.3: 配列上の類似部分の抽出

3.3節で述べたように、本手法は多義語wを含む文から単語の配列を抽出し、得られた配列の集合をwの文脈とする。すなわち、配列の集合はwの語義選択のための手がかりを含むものと考える。今、多義語wが語義s_iとして現れている文を$S_1, S_2, \cdots, S_l$とし、任意の$S_i(1 \leq i \leq l)$から得られる配列の集合をQ_{S_i}で表す。配列の集合が手がかりを含むということは、$Q_{S_1}, Q_{S_2}, \cdots, Q_{S_l}$の各々に$w$を$s_i$に導くための手がかりが現れることを意味する。$Q_{S_1}, Q_{S_2}, \cdots, Q_{S_l}$は同一の語義$s_i$に対する文脈表現であるため、実際に現れる手がかりには共通性が見られる。

例として、動詞 “fire” が語義 “go off or discharge” として現れている文から、本研究の予備調査として獲得した配列の集合の一部を図3.3に示す。図3.3では、得られた配列の集合に斜体で示す類似部分が共通して現れていることが観察できる。共通部分は “fire” が目的語に銃や弾丸などをとることを表しており、このことは語義 “go off or discharge” の手がかりとして理に適っている。

本研究では、このように配列上に現れた手がかりを配列パターン (sequence pattern) と呼ぶ。配列パターンは順序集合$p = (x_1, x_2, \cdots, x_N)$として表す。こ

(i) (fire,OBJ,weapon,the?)

(ii) (bet,that?,.*,[SUB|be])

(iii) (drug)

図 3.4: 配列パターンの例

こで、要素x_iは次のいずれかの値をとる。

$$x_i = \begin{cases} w & \text{単語}w\text{が対応することを示す。} \\ [w_1|w_2|\cdots|w_n] & \text{単語}w_1, w_2, \cdots, w_n\text{のいずれか一つが対応することを示す。} \\ . & \text{任意の単語が対応することを示す。} \\ x* & \text{要素}x\text{が0個以上連続して対応することを示す。} \\ x? & \text{要素}x\text{が対応するが、対応しない場合にもペナルティが与えられないことを示す。} \end{cases}$$

図3.4に配列パターンの例を示す。(i) は図3.3で観察された類似部分から獲得したパターンである。末尾の要素は、単語 “the” に対応しない場合にもペナルティが与えられないことを示している。(ii) は3.5節で述べる多義性解消の実験において、訓練データから実際に獲得したパターンである。2番目の要素は “that?” であり、that 節における that の省略に対応する。3番目の要素は任意の要素が複数連続して現れることを示す。末尾の要素は “SUB” か “be” のどちらかに対応するものである。(iii) は要素数が1のパターンであり、単語の配列の任意の位置に “drug” が現れることを意味するパターンである。すなわち、本研究におけるアライメントアルゴリズムでは、アライメント結果の先頭と末尾に位置するギャップにペナルティを与えない (付録参照)。

本アルゴリズムでは、Step 1で訓練データから語義ごとに配列パターンの集合$P_1, P_2, \cdots, P_n$を獲得する。獲得の基本的な方針は、訓練データから特定の語義を含む文を多数収集し、文ごとに単語の配列の集合を求め、配列上の類似部分を観察することである。図3.5に “fire” の語義 “go off or discharge” と “terminate the employment” に対して獲得した配列パターンの集合を示す。現段階では、配列パターンの獲得は人手で行っている。

go off or discharge

P_1 p_{11} : (fire,SUB,person)
p_{12} : ([fire|set_up],OBJ,[weapon|rocket])
p_{13} : (fire,[on|upon|at],physical_object)
p_{14} : (load,[into|with],weapon)

terminate the employment

P_2 p_{21} : (fire,[SUB|by],company)
p_{22} : (fire,OBJ,[person|people|staff])
p_{23} : (fire,from,organization)
p_{24} : (hire)
p_{25} : (job)

図 3.5: “fire” の語義に対する配列パターンの集合

3.4.2 文脈の類似度の算出

Step 4では、Step 1で獲得した配列パターンの集合$P_1, P_2, \cdots, P_n$をそれぞれ語義$s_1, s_2, \cdots, s_n$の文脈情報と見なし、各パターン集合と多義語の文脈Q_{S_w}との類似性を評価する。式(3.1)で定義する$sim(P_i, Q_{S_w})$はP_iとQ_{S_w}の類似度を表しており、多義語の文脈Q_{S_w}がパターン集合P_iとどの程度適合するかを量的に求めるものである。

$$sim(P_i, Q_{S_w}) = \sum_{p_j \in P_i} (a_j + \max_{q_k \in Q_{S_w}} AS(p_j, q_k)) \tag{3.1}$$

ここで$AS(p_j, q_k)$は定義3. 6、3. 7、3. 8に従って求められる。式(3.1)ではp_jは配列パターン$(x_1, x_2, \cdots, x_N)$を表しているため、要素x_iが単語でない場合に用いる評価値$d(x_i, w)$について、次のように定義する。

$$d([w_1|w_2|\cdots|w_n], w) = \max_{1 \le i \le n} d(w, w_i) \tag{3.2}$$

$$d(\text{“.”}, w) = 0 \tag{3.3}$$

式(3.3)は任意の要素との対応を表す“.”が、任意の単語と評価値0で対応することを示している。評価値が$-1 \le d(w, w') \le 1$の範囲の値をとることから、中立的な評価値として0を与えている。

式(3.1)におけるa_jは配列パターンp_jに固有の重みを意味しており、次の式

Q_{S_1} : (think,is,to,keep_away,and,try,then)
(think,is,to,keep_away,and,try,to,find,OBJ,way,a)
(think,is,to,keep_away,and,try,to,find,OBJ,
way,to,*bury,OBJ,hatchet,the*)
...

Q_{S_2} : (seem,to,buried,all_but)
(seem,to,*buried,OBJ,hatchet,the*)
(seem,to,buried,and,preparing,for,relationship,new)
...

Q_{S_3} : (be,wrong,to,welcome,not)
(be,wrong,to,welcome,OBJ,decision,the)
(be,wrong,to,welcome,OBJ,decision,to,*bury,OBJ,hatchet,the*)
...

図 3.6: 慣用句 "bury the hatchet" を含む配列の集合

で定義する。

$$a_j = \begin{cases} u_j & \text{if } \max\limits_{q_k \in Q_{S_w}} AS(p_j, q_k) \geq t_j \\ v_j & \text{otherwise} \end{cases} \tag{3.4}$$

u_j、v_jはパターンp_j固有の定数を表しており、t_jはp_j固有の閾値を表している。これらの値の設定については、本章3.5節で言及する手法の評価実験の際には配列パターンの生成作業の一環として人手で行った。これらの値を訓練データからの統計情報に基づいて自動的に設定する試みについては、第4章に詳述する。

重みa_jは、特定の配列パターンが単独で語義決定に大きく影響する場合に対応するものとして導入するものである。例えば、多義語が慣用句を構成する語の一つである場合を考える。慣用句を含む文をコーパスから収集して配列を並べてみると、完全に一致した共通の部分配列が観察される。図3.6に示した配列群は慣用句 "bury the hatchet(和睦する)" を含む文から抽出されたものの一部である。ここでは部分配列(bury,OBJ,hatchet,the)が完全に一致した形で各集合に共通に含まれている様子が観察される。このとき、配列パターンとしてp = (bury,OBJ,hatchet,the)を与えると、該当する部分配列を含む配列qとpとのアライメントスコアは、pの要素数が4であることから$AS(p, q) = 4$

となる (定義3. 6 、3. 8 参照)。逆にpとのアライメントスコアが$AS(p, q') = 4$となる配列q'には部分配列 (`bury,OBJ,hatchet,the`) が含まれていることとなり、その配列が抽出された文には慣用句 “bury the hatchet” が現れていることとなる。このとき、閾値として$t = 4$を設定し、uに正の値、vに負の値を設定することによって、配列パターンpとの適合結果に大きな差をつける。

このほかの例として、特定の語義が自動詞 (intransitive verb) だけに対応する場合を考える。動詞 “bother” の語義 “take the trouble to do something” は、自動詞に対応する。配列パターン$p =$ (`bother,OBJ`) を考えたとき、$AS(p, q) = 2$を満足する配列qには “bother” が他動詞として現れていることとなる。すなわち、語義 “take the trouble to do something” は解の候補から外して処理を進めることができる。このような場合には、閾値$t = 2$を設定し、uに負の値を設定することによって、“bother” が目的格を持つパターンpとの適合にペナルティを与える。

式 (3.1) から、多義語の文脈との類似度が最大となるパターン集合P_{max}を求めることができる。本手法で選択する語義は、P_{max}に対応する語義s_{max}である。

3.4.3 手法の適用例

図3.7に本手法を用いて動詞 “fire” の多義性を解消する例を示す。入力文S_wは “Police said Haga was immediately fired from the force.” である。Step 1では、図3.5の配列パターンの集合P_1、P_2が獲得されたものとする。Step 2、Step 3では、S_wから単語の配列の集合$Q_{S_w} = \{q_1, q_2, q_3, q_4, q_5\}$が求められる (図3.1参照)。Step 4では$P_1$、$P_2$と$Q_{S_w}$との類似度が式 (3.1) を用いてそれぞれ求められる。式 (3.1) では、各パターンごとにQ_{S_w}の各要素とのアライメントスコア$AS(p, q)$が求められる。図3.7の下段の表はパターンごとの最大のアライメントスコアと、そのときのQ_{S_w}の要素を示している。単純のため、重みa_jをすべてのパターンに対して$u_j = v_j = t_j = 0$と定義すると、文脈の類似度は単に各アライメントスコアの総和として求められる。最終的に類似度の大きなパターン集合P_2に対応する語義 “terminate the employment” が、S_wにおける “fire” の語義として選択される。

S_w : “Police said Haga was immediately fired from the force.”

q_1 : `(said,SUB,police)`

q_2 : `(said,fired,OBJ,Haga)`

q_3 : `(said,fired,was)`

q_4 : `(said,fired,immediately)`

q_5 : `(said,fired,from,force,the)`

P_1	$\arg\max_{q_k} AS(p_j, q_k)$	$\max_{q_k} AS(p_j, q_k)$
p_{11}	q_2	1.0000
p_{12}	q_2	2.7550
p_{13}	q_2	0.8200
p_{14}	q_2	-1.8637
$sim(P_1, Q_{S_w}) = 2.7113$		

P_2	$\arg\max_{q_k} AS(p_j, q_k)$	$\max_{q_k} AS(p_j, q_k)$
p_{21}	q_5	0.9592
p_{22}	q_2	3.0000
p_{23}	q_5	2.9688
p_{24}	-	-1.0000
p_{25}	q_2	0.9592
$sim(P_2, Q_{S_w}) = 6.8871$		

図 3.7: 本手法の適用例

3.5 実験

多義性解消の作業では、各多義語の各語義ごとに手がかりとなる情報を与えるため、一般にその実験には大規模なコーパスが必要となる。特に教師付き学習に基づいた手法の場合には、その精度はコーパスの規模と品質に影響されるものと考えられる[14]。従って、精度の観点から単純に手法を評価することは難しい。

こうした問題を背景に開催された多義性解消のコンテストがSENSEVALである。SENSEVALは過去に二度開催されている。初回のSENSEVALは1998年に開催された。第二回のSENSEVALは対象言語を増やし、SENSEVAL-2という名称で2001年に開催されている。これに伴ない、初回のSENSEVALはしばしばSENSEVAL-1という名称で参照されている。本論文では以降、混乱を起こさないために初回のSENSEVALをSENSEVAL-1という。本節は、SENSEVAL-1の動詞を対象に行った多義性解消の実験とその結果について述べるものである。

本実験は3.4節で述べた手順に沿って動詞の多義性解消を行うものである。まず、訓練データとしてSENSEVAL-1のトレーニングコーパスを用い、そこから得られる配列だけに基づいて人手で配列パターンを作成する。次に、実験データとしてSENSEVAL-1のテストコーパスを用い、各多義語に式(3.1)に基づいて推定される語義を割り当てる。本実験では、式(3.4)で用いられているu_j、v_j、t_jとして、配列パターンの生成作業の一貫として人手で付与したものを用いている。また、Step 2の構文解析にはApple Pie Parser[39]を用いた。配列の抽出には、Apple Pie Parserから得られた解析結果にノード“SUB”、“OBJ”を追加し、誤りを人手で修正したものを用いている。

表3.1に実験結果を示す。ここではSENSEVAL-1におけるシステムの評価[40]と同様に、語義の粒度に従ってfine-grained、mixed-grained、corase-grainedの三つの観点で適合率と再現率を求めている。表の各欄の値は、それぞれ適合率/再現率の組を示しており、括弧内の値はSENSEVAL-1に参加したシステムが各単語に対して達成した精度の中で最良のものを示している[4]。また、図3.8に本実験で作成した配列パターンのうち、“bet”に対するものを示す。各配列パターンの末尾にある三つ組は、そのパターンに対する重み関数a_jを構成するものとして人手で付与したu_j、v_j、t_jの組を表している。

[4]括弧内の精度は`http://www.senseval.org/`より引用したものである。

表 3.1: SENSEVAL-1の動詞に対する実験結果

対象動詞	試行数	fine-grained	mixed-grained	coarse-grained
amaze	70	1.000/1.000 (1.000/1.000)	1.000/1.000 (1.000/1.000)	1.000/1.000 (1.000/1.000)
bet	117	0.880/0.880 (0.778/0.778)	0.897/0.897 (0.786/0.786)	0.906/0.906 (0.838/0.838)
bother	209	0.900/0.900 (0.866/0.866)	0.900/0.900 (0.880/0.880)	0.900/0.900 (0.880/0.880)
bury	201	0.667/0.667 (0.572/0.572)	0.678/0.678 (0.578/0.578)	0.682/0.682 (0.592/0.592)
calculate	218	0.950/0.950 (0.922/0.922)	0.950/0.950 (0.922/0.922)	0.950/0.950 (0.922/0.922)
consume	186	0.645/0.645 (0.503/0.500)	0.704/0.704 (0.586/0.583)	0.726/0.726 (0.616/0.613)
derive	217	0.751/0.751 (0.664/0.664)	0.758/0.758 (0.677/0.677)	0.760/0.760 (0.687/0.687)
float	229	0.616/0.616 (0.555/0.555)	0.655/0.655 (0.614/0.614)	0.672/0.672 (0.629/0.629)
invade	207	0.686/0.686 (0.556/0.556)	0.766/0.766 (0.623/0.623)	0.778/0.778 (0.662/0.662)
promise	224	0.942/0.942 (0.906/0.906)	0.942/0.942 (0.911/0.911)	0.942/0.942 (0.911/0.911)
sack	178	0.989/0.989 (0.978/0.978)	0.989/0.989 (0.978/0.978)	0.989/0.989 (0.978/0.978)
scrap	186	0.935/0.935 (0.898/0.898)	0.978/0.978 (0.978/0.978)	0.978/0.978 (0.978/0.978)
seize	259	0.768/0.768 (0.714/0.714)	0.768/0.768 (0.714/0.714)	0.776/0.776 (0.753/0.753)
all items	2501	0.811/0.811 (0.709/0.709)	0.831/0.831 (0.742/0.741)	0.838/0.838 (0.755/0.755)

各欄の値は“適合率/再現率”を表す。また、括弧内の値は各々の対象動詞に対してSENSEVAL-1参加システムが達成した最良の値を示す。

gamble (519907)

(bet,[on|against],[horse|jockey|race]) (0.00, 0.00, 0.00)
(bet,OBJ,possession) (−5.00, 5.00, 0.00)
(bookmaker) (0.00, 0.00, 0.00)
(ladbrokes) (0.00, 0.00, 0.00)
(bet,with,possession) (0.00, 0.00, 0.00)

money (519916)

(bet,OBJ,possession) (0.00, −5.00, 1.80)
(bet,[on|against],[horse|jockey|race|racing]) (0.00, 0.00, 0.00)
(bookmaker) (0.00, 0.00, 0.00)
(ladbrokes) (0.00, 0.00, 0.00)

ditrans (521071)

(bet,OBJ,person) (0.00, −5.00, 1.80)
(bet,OBJ,possession) (0.00, 0.00, 0.00)
(bet,that?,.*,[SUB|be]) (0.00, −5.00, 1.60)
(bet,[on|against],[horse|jockey|race|racing]) (−5.00, 5.00, 0.00)

think (519908)

(bet,will) (0.00, 0.00, 0.00)
(mind,betting) (0.00, 0.00, 0.00)
(bet,OBJ) (−4.00, 4.00, 0.00)
(bet,SUB,I) (0.00, −5.00, 2.40)
(bet,that?,.*,[SUB|be]) (0.00, −1.00, 1.60)

speculate (520051)

(bet,[on|against],[act|attainment|status|business]) (0.00, 0.00, 0.00)
(bet,OBJ,possession) (−5.00, 5.00, 0.00)
(bet,that?,.*,[SUB|be]) (0.00, −1.00, 1.60)
(speculator) (0.00, 0.00, 0.00)
(rumor_monger) (0.00, 0.00, 0.00)

assume (520048)^a

(bet,that?,.,be) (0.00, −1.00, 1.60)
(bet,OBJ,[life|ass],your) (0.00, −6.00, 2.90)
(bet,that?,.,SUB) (0.00, −1.00, 1.80)

assume (520048)^b

(bet,SUB,you) (0.00, −5.00, 2.40)
(bet,[can|be_willing_to]) (0.00, −4.00, 1.60)
(bet,that?,.,be) (0.00, −1.00, 1.60)
(bet,[that],.,SUB) (0.00, −1.00, 1.60)

unlikely (520050)

(bet,SUB,I) (0.00, −5.00, 2.40)
(bet,on,it) (0.00, −5.00, 1.60)
(bet,[would_not|do_not]) (0.00, −5.00, 1.60)

図 3.8: 動詞 “bet” の配列パターン

3.6 考察

本手法は人手による教師付き学習を行っているため、SENSEVAL-1の参加システムとの精度の単純比較による評価は困難である。しかし、表3.1によると、本手法を適用することによってすべての対象の精度が向上している[5]。精度向上の程度には評価方法や参加システムの精度に応じたばらつきはあるが、語義の粒度が最も細かいfine-grained scoringでは、最大で14.2%の精度向上が見られる。こうした精度改善は本手法を多義性解消に適用することの妥当性を示唆するものと評価できる。

表3.2に人手による各単語の多義性解消の精度を示す。表3.1と表3.2によると、本手法は“calculate”、“promise”、“sack”などに関して人手による多義性解消に近い精度が達成できている。一方、“bury”、“consume”、“float”、“invade”などに関しては、人手による精度に対して劣る結果となっている。ここで、“calculate”、“promise”、“sack”のfine-grainedの語義に対するentropyはそれぞれ0.982、0.982、0.132であるのに対し、“bury”、“consume”、“float”、“invade”はそれぞれ2.759、2.218、3.333、2.195である[34]。ここで用いられているentropyは語義の出現のばらつきを表すものであり、従ってentropyの高い後者の単語群に関しては多義性解消が困難な傾向が認められる。得られた精度のばらつきには、こうした傾向が影響していると考えられる。

また、いくつかの単語に対する精度が人手による精度に劣っていることについては、定義3. 6における評価値$d(w, w')$が最も大きく影響しているものと推測している。Semantic DistanceはWordNetにおけるノード間の距離を推定するものであるため、定義3. 6では二つの単語の語義すべての組み合わせのうち、最小の距離を持つ語義を用いてスコアを求めている。この方法はしばしば不適切な値を導く。例えば、配列$q =$ `(fire,OBJ,gun,a)`を考える。図3.5の配列パターンを用いるとき、この配列はパターン$p =$ `([fire|set_up],OBJ,[weapon|rocket])`に対して最良のアライメントスコアを得ることが期待される。しかし“gun”は“a professional killer who uses a gun”の意味も持っているため、パターン$p' =$ `(fire,OBJ,[person|people|staff])`に対しても高いアライメントスコアが得られてしまう。両パターンに対するアライメントスコアの差は非常に小さく、$AS(p, q) - AS(p', q) = 0.115$である。これは単語間の類似度を表す評価

[5]“amaze”は動詞の語義を一つしか持っておらず、ここでは考察の対象から除いている。

表 3.2: 人手による動詞の多義性解消の精度

対象動詞	fine-grained	mixed-grained	coarse-grained
amaze	1.000/1.000	1.000/1.000	1.000/1.000
bet	0.924/0.916	0.932/0.925	0.932/0.925
bother	0.976/0.976	0.976/0.976	0.976/0.976
bury	0.928/0.923	0.930/0.925	0.933/0.928
calculate	0.954/0.950	0.959/0.954	0.959/0.954
consume	0.944/0.939	0.955/0.950	0.958/0.953
derive	0.965/0.961	0.965/0.961	0.965/0.961
float	0.927/0.923	0.938/0.934	0.943/0.939
invade	0.921/0.912	0.922/0.913	0.924/0.915
promise	0.953/0.953	0.962/0.962	0.962/0.962
sack	0.994/0.994	0.994/0.994	0.994/0.994
scrap	0.981/0.981	0.995/0.995	0.995/0.995
seize	0.921/0.921	0.921/0.921	0.929/0.929
all items	0.950/0.947	0.955/0.952	0.957/0.954

各欄の値は “適合率/再現率” を表す。

値$d(w, w')$がそれぞれ

$$d(\text{“gun”}, \text{“weapon”}) = 0.990$$
$$d(\text{“gun”}, \text{“person”}) = 0.875$$

であることが原因である。

より適切なアライメントスコアを得るためには、$d(w, w')$の定義式に対し、単語間の類似度をより直観的な値として求められるよう改良を施す必要がある。単語間の妥当な類似度の推定に関しては既に多くの試み[41, 42, 43, 44]がなされており、より妥当な類似度を定義することによって、本手法による多義性解消の精度はさらに向上するものと期待される。

3.7 まとめ

本章では、本研究で構築した単語の配列に基づく多義性解消の新しい手法について詳述した。本手法で用いられる単語の配列は文の依存構造木全体から求められるものであり、多義語に関する連想関係、選択制限を両者とも考慮できる。すなわち、本手法は連想関係を用いた手法の特長と選択制限を用いた手法の特長を併せ持った手法である。また、ペアワイズアライメントに基づいて文脈の類似性評価を行っていることから、単純なマッチングに基づいた手法に比べ、頑健で柔軟な処理が可能である。SENSEVAL-1の動詞を対象とした多義性解消に関する実験では、本手法を用いることによってSENSEVAL-1への参加システムの最良の精度に比べて最大で14.2%の精度向上が見られた。これは本手法の妥当性が有意であることを示唆している。

本手法の問題点は大きく二つ挙げられる。一つは構文解析の問題であり、もう一つは人手による知識獲得の問題である。

前者は本手法が多義性解消の際に正しい依存構造木が利用可能と仮定している点である。しかし構文解析システムの精度は年々改良されてきており、この問題は構文解析技術の発展に伴ってますます小さな問題になっていくものと考えられる。また本手法は依存構造から得られる配列に対し、その類似性を数値として求めているため、確率的構文解析システムとの統合が比較的容易なものと考えられる。

後者の問題は、語義選択の手がかりである配列パターンと、各パターンごとに定義される重みとを人手によって獲得している点である。機械翻訳システ

ムや情報検索システムなど、実際のシステムに本手法を実装する際には、こうした語義知識の獲得コストを軽減することが手法実装のコストを軽減することに直接結びつく。こうした観点に基づき、第4章では本手法の自動化を目標として、アライメントスコアに付与する最適な重みを訓練データから推定する手法について詳述する。

第 4 章

アライメントスコアの重みの推定

第3章では、本研究で構築した多義性解消の新しい手法について述べた。本手法を用いた実験では平均81.1%の高い精度で多義性解消が達成できている。しかし、本手法には人手による知識獲得が必要であり、このときの人手のコストを軽減することが問題である。ペアワイズアライメントを用いた多義性解消の自動化を目標として、本章ではアライメントスコアに付与する最適な重みを訓練データから推定する手法について詳述する。また、本手法による重みの推定を用いた動詞の多義性解消に関する実験についても言及する。実験から得られた結果は本手法の有効性を示すものであった。なお、本章で用いる単語の配列やペアワイズアライメント、およびそれに付随するいくつかの概念については、3.2節における定義に従うものとする。

4.1 はじめに

第3章は、多義性解消にペアワイズアライメントの技法を適用した新しい手法について詳説するものであった。本手法は連想関係に基づく手法の特長と選択制限に基づく手法の特長とを併せ持った手法であり、SENSEVAL-1の動詞を対象とした実験では、平均81.1%の精度で多義性解消が達成できている。

しかし、本手法には語義に関する知識獲得や入力文の構文解析などの人手による調整が介在しており、すなわち人手のコストの問題が存在する。1.1節で言及したように、本研究で構築した多義性解消の手法を応用するものとしては機械翻訳システムや情報検索システムなどが考えられる。こうしたシステムに本手法を応用するためには、あらかじめ語義知識をシステムの辞書情報として与えておく必要がある。従って本手法のシステムへの実装において、人手のコストが大きな障壁となることが考えられる。人手による語義知識獲

得のコストを軽減することは、本手法の解決すべき課題である。

第3章に述べた手法において、獲得すべき語義知識は次の二つである。

- 語義ごとに与える配列パターン
- 各配列パターンに与える重み

これらの情報を大規模コーパスから自動的に獲得することができれば、手法の実装のコストは大きく軽減されることになる。本章はペアワイズアライメントを用いた多義性解消の自動化を目標として、これらの情報のうち、配列パターンに与える重みを訓練データから自動的に推定する試みについて詳述するものである。語義の選択に対する配列パターンの振る舞いは、配列パターンに付与される重みに大きく影響される。従って重みの妥当性は、配列パターンを手がかりとした多義性解消において、精度の向上の観点から極めて重要な位置を占める。

本章で述べる重み推定の手法は、訓練データにおける統計情報に従って妥当な重みを推定するものである。本手法は重み推定を閾値の推定と重みの推定の二つの問題に分割し、前者に対して帰納的な手法を適用し、後者に対してエントロピーに基づく手法を適用する。また本章では、第3章と同様にSENSEVAL-1の動詞を対象に行った実験についても言及する。本実験では、推定された重みを用いたときの多義性解消の精度が人手による重みを用いたときの精度を下回る結果が得られたが、いくつかの単語に対してはSENSEVAL-1における最良の精度を大きく上回る精度が得られた。

本章の構成は次の通りである。4.2節では訓練データから配列パターンごとに最適な重みを推定する手法について、基本的な考え方を述べる。本手法の詳細は4.3節と4.4節に述べる。4.5節では本手法を用いて行った多義性解消の実験について述べる。最後に4.6節でまとめを述べる。

4.2 基本的な考え方

第3章で述べた多義性解消の手法の欠点の一つとして、配列パターンに対する重みの定義を人手によって行っていることが挙げられる。本節では、式(3.4)で定義される重みa_jについて、配列パターンごとに最適な定数u_j、v_j、t_jを統計的に推定する手法の概略を述べる。本手法を用いることにより、重みの定義にかかる人手のコストを大きく軽減することができる。

式(3.1)において、重みa_jは語義選択に対する各配列パターンの貢献の度合いを表すものである。3.4.2節で述べたように、例えば慣用表現などでは入力文における特定の配列パターンの有無が語義選択に大きく影響する。また、自動詞だけに対応する語義の選択は、目的語を伴う配列パターンの有無に影響される。重みを付与することによって、各配列パターンのこうした語義選択への貢献を評価値に反映させることができる。すなわち、特定の配列パターンに良く適合する配列が入力文に存在するか否かに従って、評価値に重みやペナルティを与えることが可能である。

ここで問題となるのは、配列パターンと入力文における配列との適合の判断をどのようにするかということと、どの程度の重みを付与するかということの二つである。すなわち配列パターンに重みを付与する問題は、式(3.1)における表記に従って次の二つの問題に分割することができる。

- 閾値t_jの推定
- 定数u_j、v_jの推定

推定方法に対する本手法の基本的な考えは、訓練データから最適な閾値t_jを帰納的に推定し、推定されたt_jを用いて求められる統計量から定数u_j、v_jを求めることである。以下、この二つの問題に対する手法について、それぞれ4.3節、4.4節で詳述する。

4.3 閾値の推定

式(3.4)における$\max_{q_k \in Q_{S_w}} AS(p_j, q_k)$は、任意の配列パターン$p_j$と入力された配列とのアライメントスコアのうち、最も高い値を得るものである。アライメントスコアは配列同士の適合の度合いを示す値である。従ってこの値は、p_jと最も良く適合する入力文中の配列の適合の度合いを示している。式(3.4)の条件部は、この値とp_j固有の閾値t_jとを比較することによって、p_jと入力文における配列とが適合しているかどうか判断することを意味している。

ここで、適合の判断をするための最適な閾値をどのように求めるかが問題となる。以下では訓練データにおける統計情報に従って、最適な閾値を帰納的に推定する方法について述べる。

多義語wの任意の語義をsとする。sに対して与えられた配列パターンのうち任意の一つをp_jで表す。また、訓練データのうちwを含む文を$S = \{S_1, S_2, \cdots$

$, S_N\}$とし、Sにおいてwが語義sとして現れている文を集合S'で表す。すなわち、S'はSの部分集合である。wがs以外の語義として現れている文は集合$\tilde{S}'$で表す。今、p_jとS_iにおける単語の配列とのペアワイズアライメントを用いてSを次の二つの集合に分割する。

$$T = \{S_i | \max_{q_k \in Q_{S_i}} AS(p_j, q_k) \geq t_j\} \tag{4.1}$$

$$\tilde{T} = \{S_i | \max_{q_k \in Q_{S_i}} AS(p_j, q_k) < t_j\} \tag{4.2}$$

ここで、t_jを用いてp_jの適合判断をすることの有効性を評価するため、次のように適合率(precision)と再現率(recall)を求める。

$$\text{precision}(S', T) = \frac{|S' \cap T|}{|T|} \tag{4.3}$$

$$\text{recall}(S', T) = \frac{|S' \cap T|}{|S'|} \tag{4.4}$$

式(4.3)は、p_jに適合した配列(評価値がt_jを超えた配列)が含まれる文のうち、wが正しい語義sで用いられている文の割合を示している。式(4.4)は、wが正しい語義で用いられている文のうち、p_jに適合した配列が含まれる文の割合を示している。

式(4.3)、(4.4)は、p_jに適合する配列が入力文に含まれる場合に語義sを選択することの適合率と再現率を示すものである。すなわち、これらの値が高いほどp_jとt_jを用いて語義sを選択することが妥当と見なすことができる。最も妥当な適合率と再現率の組を求めるため、F-measureを用いてこの妥当性の評価値とする。F-measureは次の式で定義される[10]。

$$F = \frac{1}{\alpha \cdot \frac{1}{\text{precision}} + (1 - \alpha) \cdot \frac{1}{\text{recall}}} \tag{4.5}$$

αは適合率と再現率に対する重み(定数)である。本手法では適合率と再現率を均等に考えるため、$\alpha = 0.5$としている。

また、閾値t_jは、p_jに適合する配列が入力文に含まれない場合に語義sを選択しないという観点でも用いられる。すなわち、p_jとt_jを用いて語義sを選択しないことの妥当性も評価する必要がある。この観点からは、次の適合率と再現率が求められる。

$$\text{precision}(\tilde{S}', \tilde{T}) = \frac{|\tilde{S}' \cap \tilde{T}|}{|\tilde{T}|} \tag{4.6}$$

$$\text{recall}(\tilde{S}', \tilde{T}) = \frac{|\tilde{S}' \cap \tilde{T}|}{|\tilde{S}'|} \tag{4.7}$$

式(4.6)、(4.7)の適合率と再現率を用いて得られるF-measureを$\tilde{F}$で表すとする。本手法では、閾値t_jの妥当性を表す評価値Rを次の式で定義する。

$$R = F + \tilde{F} \tag{4.8}$$

$0 \leq F \leq 1$、$0 \leq \tilde{F} \leq 1$であるため、$0 \leq R \leq 2$である。

一方、前述の自動詞だけに対応する語義の選択の例などでは、p_jに適合する配列が入力文に含まれる場合に語義sを選択しない、あるいは、入力文に含まれない場合にsを選択するという観点も必要となる。この際の適合率と再現率は次のように二組求められる。

$$\text{precision}(S', \tilde{T}) = \frac{|S' \cap \tilde{T}|}{|\tilde{T}|} \tag{4.9}$$

$$\text{recall}(S', \tilde{T}) = \frac{|S' \cap \tilde{T}|}{|S'|} \tag{4.10}$$

$$\text{precision}(\tilde{S'}, T) = \frac{|\tilde{S'} \cap T|}{|T|} \tag{4.11}$$

$$\text{recall}(\tilde{S'}, T) = \frac{|\tilde{S'} \cap T|}{|\tilde{S'}|} \tag{4.12}$$

これらの適合率、再現率から求められる評価値を$R' = F' + \tilde{F}'$とする。

p_jの最適な閾値t_jはRとR'の二つの評価値を基に求めることができる。すなわち、t_jの値を変化させてその都度RとR'を求め、最も高い評価値が得られたときのt_jを最適な閾値とする帰納的なアプローチで推定できる。ここで、最大の評価値Rを$\max_{t_j} R$、最大の評価値R'を$\max_{t_j} R'$で表すとすると、推定される閾値は次のように表すことができる。

$$t_j = \begin{cases} \arg\max\limits_{t_j} R, & \text{if } \max\limits_{t_j} R \geq \max\limits_{t_j} R' \\ \arg\max\limits_{t_j} R', & \text{if } \max\limits_{t_j} R < \max\limits_{t_j} R' \end{cases} \tag{4.13}$$

今、各文から得られる単語の配列のうち最もp_jに適合する配列q_kを取り上げ、このときのp_jとq_kの適合の度合い$AS(p_j, q_k)$を考える。$AS(p_j, q_k)$が閾値t_jを超えればパターンp_jに適合した配列q_kが含まれることになる。逆に$AS(p_j, q_k)$がt_jを超えなければ、その文から得られた単語の配列のうち最もp_jに適合するq_kが基準を満たさないという観点で、p_jに適合した配列は含まれないことになる。t_jはこうした判断の基準となる閾値であり、式(4.13)はこの閾値t_jを求めるものである。

4.4 重みの推定

4.3節で定義した式(4.13)は、訓練データにおける各文においてパターンp_jに適合した配列が含まれるかどうかの基準を示す閾値である。閾値を推定することによって、特定の配列パターンに良く適合する配列が入力文に存在するか否かの判断が可能となる。ここで残された問題は、アライメントスコアに基づく評価値$\max_{q_k \in Q_{S_w}} AS(p_j, q_k)$に対し、こうした判断にしたがってどの程度の重みやペナルティを付与するかという問題である。本小節では、訓練データからの統計情報に従って各配列パターンの語義選択に対する影響の度合いを定義し、これを基に式(3.4)の定数u_j、v_jを定義する。

多義語wの語義を$s_1, s_2, \cdots, s_n$とし、訓練データのうちwを含む文の集合を$S = \{S_1, S_2, \cdots, S_N\}$とする。任意の$S_i$に配列パターン$p_j$と適合した配列が存在するときに、$S_i$における$w$の語義が$s_k$となる確率を$\Pr(s_k|t_j)$で表す。$p_j$に適合した配列が存在しないときに語義が$s_k$となる確率は$\Pr(s_k|\tilde{t_j})$で表す。また、$w$が語義$s_k$として$S_i$に現れるときに、$S_i$に配列パターン$p_j$に適合した配列が存在する確率を$\Pr(t_j|s_k)$で表し、存在しない確率を$\Pr(\tilde{t_j}|s_k)$で表す。語義$s_k$に対して与えられた配列パターンのうち任意の一つを$p_j$とすると、$p_j$の語義選択に対する影響の度合いとして次の二つのエントロピーを定義する。

$$H_{t_j}(s) = -\sum_{i=1}^{n} \Pr(s_i|t_j) \log \Pr(s_i|t_j) \tag{4.14}$$

$$H_{s_k}(t) = -\Pr(t_j|s_k) \log \Pr(t_j|s_k) - \Pr(\tilde{t_j}|s_i) \log \Pr(\tilde{t_j}|s_i) \tag{4.15}$$

$H_{t_j}(s)$はp_jに適合する配列が任意のS_iに存在するときに、S_iにおけるwの語義がどの程度ばらついているかを示す量である。この値が低いほど語義のばらつきは小さく、従ってp_jが語義の候補を絞り込んでいる傾向が強いと考えることができる。一方、$H_{s_k}(t)$はwが語義s_kとして現れている文に、p_jに適合する配列が存在する傾向についてのばらつきを示す。この値が低いほどp_jに適合する配列の出現と語義s_kの出現との間には関連があり、従ってp_jとs_kの結び付きが強いと考えることができる。

本手法ではこれらの尺度を用いて、$t_j = \arg\max_{t_j} R$のときの重み定数u_j、v_jを次のように定義する。ここで、配列パターンp_jは語義s_kに対して与えられたものであるとし、また$H_{t_j}(s) \neq 0$、$H_{s_k}(t) \neq 0$とする。

$$u_j = \frac{\Pr(s_k|t_j)}{H_{t_j}(s)} \cdot |p_j| \tag{4.16}$$

$$v_j = -\frac{\Pr(t_j|s_k)}{H_{s_k}(t)} \cdot |p_j| \tag{4.17}$$

ここで、$|p_j|$は配列パターンp_jの要素数(長さ)を表す。また、訓練データのデータスパースネス(data sparseness)の問題から$H_{t_j}(s) = 0$となるときには$u_j = |p_j|$とし、$H_{s_k}(t) = 0$となるときには$v_j = -|p_j|$とする。

式(4.16)において、$\frac{1}{H_{t_j}(s)}$は配列パターンp_jが語義の候補を絞り込むほど大きな値を取るが、この値は適切な語義s_kに絞り込むことを保証していないため、$\Pr(s_k|t_j)$との積を求めている。従って式(4.16)は、配列パターンp_jが多義語wの語義候補を適切な語義s_kに絞り込むほど大きな値を取ることになる。

一方、式(4.17)における$\frac{1}{H_{s_k}(t)}$は、語義s_kの出現と配列パターンp_jの適合との関連性が強いほど大きな値を取る。ここでの関連性とは「語義s_kが文に含まれるときは、その文に配列パターンp_jに適合する配列が含まれる」という言わば正の関連性と、「語義s_kが文に含まれるときは、その文に配列パターンp_jに適合する配列が存在しない」という言わば負の関連性とがある。式(4.17)の定義では前者の関連性を期待しているが、$\frac{1}{H_{s_k}(t)}$は関連性の正負を区別しないため、$\Pr(t_j|s_k)$との積を求めている。従って式(4.17)は、語義s_kと配列パターンp_jとの正の関連性が強いほど、負の方向に大きな値を取ることになる。

一方$t_j = \arg\max_{t_j} R'$のとき、重みは$t_j = \arg\max_{t_j} R$のときと逆の観点から次のように定義する。

$$u_j = -\frac{\Pr(t_j|s_k)}{H_{s_k}(t)} \cdot |p_j| \tag{4.18}$$

$$v_j = \frac{\Pr(s_k|t_j)}{H_{\tilde{t_j}}(s)} \cdot |p_j| \tag{4.19}$$

ここで、$H_{\tilde{t_j}}(s)$は次の式で与えられる。

$$H_{\tilde{t_j}}(s) = -\sum_{i=1}^{n} \Pr(s_i|\tilde{t_j}) \log \Pr(s_i|\tilde{t_j}) \tag{4.20}$$

同様に$H_{\tilde{t_j}}(s) = 0$となるときには$u_j = -|p_j|$とし、$H_{s_k}(t) = 0$となるときには$v_j = |p_j|$とする。

4.5 実験

本節では推定された重みを用いて行った多義性解消の実験について述べる。本実験では推定された重みの妥当性を正しく評価するため、3.5節における実験と同様に、対象をSENSEVAL-1の動詞とする。実験に用いる配列パターン

についても、3.5節の実験の際に用いたパターンと同じものを用いることとする。本実験では訓練データをSENSEVAL-1のトレーニングコーパスとする。実験の手順は、

手順1. すべての配列パターンに対して4.3節で述べた手法に基づいて閾値を推定する。

手順2. 手順1.で推定された閾値を用いて、すべての配列パターンに対して重みを推定する。

手順3. 手順1.で推定された閾値と手順2.で推定された重みを用いて、3.4節に述べた手順に従って多義性解消を行なう。

である。試験データは3.5節と同様に、SENSEVAL-1のテストコーパスである。

ここで、手順1.の閾値の推定では式(4.13)における最良の評価値R、R'を推定するために、t_jを変化させる必要がある。t_jは式(4.1)、(4.2)における条件部$\max_{q_k \in Q_{S_w}} AS(p_j, q_k) \geq t_j$、$\max_{q_k \in Q_{S_w}} AS(p_j, q_k) < t_j$に用いるものであり、この条件に従って各適合率、再現率が求められ、評価値R、R'が求められる。このとき、定義3. 6、3. 7、3. 8と、本研究で用いるペアワイズアライメントのアルゴリズム(付録参照)から、t_jの比較対象である$\max_{q_k \in Q_{S_w}} AS(p_j, q_k)$は$-|p_j| \leq \max_{q_k \in Q_{S_w}} AS(p_j, q_k) \leq |p_j|$である。従って$t_j$の推定範囲を$-|p_j| \leq t_j \leq |p_j|$とする。また、本実験では$t_j$の推定の粒度は、計算コストなどから総合的に判断して0.01としている。

表4.1に実験結果を示す。ここではSENSEVAL-1におけるシステムの評価のうち、語義の粒度を最も細かく評価するfine-grained scoringによって精度を求めている。表の列「本手法による」は、本手法を用いて推定された重みを用いた場合の多義性解消の精度を示している。列「人手による」は人手による重みを用いたときの精度を示しており、3.5節に示した表3.1における評価値と同じ値である。列「SENSEVAL-1」は、SENSEVAL-1に参加したシステムが各単語に対して達成した精度の中で最良のものを示している。また、表の各欄の値はそれぞれ適合率/再現率の組を示している。

表4.1によると、本手法の重みを用いた多義性解消の精度が “bet”、“consume”、“derive”、“invade”、“all items” の5つの項目でSENSEVAL-1の最良の精度を上回る結果が得られた。特に “invade” では10.6%の精度向上が見られる。人手による重みを用いた場合の精度を下回るものとなってはいるが、この結果は本

表 4.1: SENSEVAL-1 の動詞に対する実験結果

対象動詞	試行数	本手法による	人手による	SENSEVAL-1
amaze	70	1.000/1.000	1.000/1.000	1.000/1.000
bet	117	0.786/0.786	0.880/0.880	0.778/0.778
bother	209	0.713/0.713	0.900/0.900	0.866/0.866
bury	201	0.463/0.463	0.667/0.667	0.572/0.572
calculate	218	0.835/0.835	0.950/0.950	0.922/0.922
consume	186	0.586/0.586	0.645/0.645	0.503/0.500
derive	217	0.751/0.751	0.751/0.751	0.664/0.664
float	229	0.485/0.485	0.616/0.616	0.555/0.555
invade	207	0.662/0.662	0.686/0.686	0.556/0.556
promise	224	0.808/0.808	0.942/0.942	0.906/0.906
sack	178	0.978/0.978	0.989/0.989	0.978/0.978
scrap	186	0.823/0.823	0.935/0.935	0.898/0.898
seize	259	0.656/0.656	0.768/0.768	0.714/0.714
all items	2501	0.713/0.713	0.811/0.811	0.709/0.709

各欄の値は “適合率/再現率” を表す。

手法の有効性を示すものである。“all items” については本手法による精度向上の幅が小さいが、この理由としては、各語義に対する知識の獲得手法に統一性が欠けていることを考えている。すなわち、配列パターンは人手によって獲得しており、重みは訓練データから統計的に推定されたものであるため、語義知識の獲得には統一性が欠けるものと考えられる。配列パターンを訓練データから自動的に獲得するアルゴリズムを構築し、機械的に獲得された配列パターンを用いることによって、本手法の精度は人手による重みを用いたときの精度に近づくものと期待される。

4.6 まとめ

本章では、アライメントスコアに付与する最適な重みを訓練データから推定する手法について詳述した。本手法は配列パターンごとに固有の重みを訓練データからの統計情報のみで推定するものであり、従って人手のコストを考えずに済む。統計情報を用いることによってデータスパースネスなどの問題が浮上する反面、客観的で信頼性の高い重みを決定することができる。

4.5節では本手法により推定された重みを用いた多義性解消の実験について言及した。本実験では推定された重みの妥当性が人手による重みの妥当性に劣る結果が得られた。しかし、推定された重みを用いた多義性解消の精度は、いくつかの単語に対してSENSEVAL-1における最良の精度を上回るものであった。すべての項目を比較した場合の精度向上の幅は小さいが、この結果は本手法による重みの推定が有効なものであることを示している。

第 5 章

文照合への応用

5.1 はじめに

本章では、本研究で構築した多義性解消の手法を文照合の問題に応用することについて検討する。

本章で扱う文照合の問題とは、与えられた二つの文を照らし合わせて両者の類似性を判断する問題であり、すなわち二つの文の類似度を求める問題として定義できる。文照合の応用範囲は広く、用例に基づく機械翻訳 (example-based MT)、情報検索、文章の自動校正などが例として挙げられる。文照合は自然言語処理における有用性の高いものの一つとして位置づけられる [45]。

文照合の最も基本的なものは、文の表層文字列間での照合である。本研究で単語の配列間の類似度を求めるために用いたペアワイズアライメントは、本来この問題を解決するものとして構築された手法である。ペアワイズアライメントの代表的な応用にDNA配列の類似性評価がある。DNA配列は四種類の塩基アデニン (adenine)、チミン (thymine)、グアニン (guanine)、シトシン (cytosine) の並びによって生物の遺伝情報を表すものである。各塩基にはそれぞれ “A”、“C”、“G”、“T” の文字が割り当てられるため、塩基の並びであるDNA配列は一種の文字列として見なされる。また、DNAの情報は結局はタンパク質に翻訳することによって発現されるが、タンパク質も20種類のアミノ酸からなる文字列として表現される。すなわち、DNA配列の類似性評価はペアワイズアライメントを用いた文字列照合の問題に帰着される [38]。

しかし、文照合のさまざまな応用において、表層文字列の水準における照合では要求される類似性判断の精度を満足する結果が得られない場面が多数存在する。例えば用例に基づく機械翻訳は入力文に類似する用例を用いて翻

訳を行うものである。このとき、入力文と用例の類似性判断には、表層格の一致する格要素(単語や句など)を比較する必要がある。すなわち、文の構文情報までを用いた照合の問題を解決しなければならない。また、情報検索では構文情報を用いることによって文照合の精度が向上し、検索精度に良い影響を与えることが予想される。このように、照合に表層文字列だけでなく文の構文情報まで用いることは、文照合の精度を向上させるための妥当な展開であると考察される。このとき、構文情報の類似性をいかにして評価するかが問題となる。

本論文ではこれまで、本研究で構築した多義性解消の手法について、第3章、第4章を通じて説明してきた。本手法は、多義性解消の基本的な手順として次のようなアプローチを用いることによって特徴づけられる。

- 多義語の文脈表現として単語の配列を用いる
- 文脈の類似性評価にペアワイズアライメントを用いる

3.3節に言及したように、単語の配列は多義語の出現する文の依存構造から抽出されるため、文の依存構造をよく保存する性質を持つ。すなわち、一文から得られた単語の配列の類似性を評価することは、文の依存構造の類似性を評価することと等価なものとして捉えることができる。

こうした観点の下、本章では本研究で構築した多義性解消の手法を文照合に応用することについて論じる。本章の構成は次の通りである。5.2節では、文の構文構造の類似度を求める問題について関連する先行研究を概説する。5.3節では、本手法を文照合問題に応用する際の検討に基づいて文照合を定式化する。5.4節では本章の検討の妥当性を確認するために行った小規模な実験について言及し、5.5節でまとめを述べる。

5.2 文照合に関連する先行研究

本研究では定義3. 2 のように、任意の文Sの依存構造を二つ組(V, E)で定義される有向木と見なしている。ここで、Vは文Sに含まれるすべての単語の集合であり、EはSに含まれるすべての係り受け関係の集合である。すなわち、本研究では文の依存構造を大局的にグラフ(graph)の一種として捉えるものである。グラフの類似性を評価する問題に対しては、さまざまな分野での応用を目的とした研究が多数報告されている。ここではこれらの先行研究のうち、

グラフ類似度の分子構造類似性解析への応用と構造的パターン認識への応用を目的としたものについて概説する。

Takahashi and Ishiyamaは分子の化学構造式をグラフとして扱い、類似した化学構造の検索システムを実現することを目的として、グラフの類似度を定義している[46]。Takahashi and Ishiyamaの手法では、グラフからすべての可能な部分グラフが列挙され、個々の部分グラフに対して定量的特徴付けが行われる。特徴付けには部分グラフを構成する頂点の次数和、頂点に対応する原子(団)の質量数の総和などが用いられ、特徴づけられた個々の部分グラフの特徴指数に従った度数分布が調べられる。この度数分布は一種の多次元パターンと見なされ、ベクトルを用いて表現される。二つのグラフの類似度は、それぞれのベクトルのユークリッド距離を基に求められる。

Bunke and Allermannは構造的パターン認識への応用を目的としたグラフマッチングを報告している[47]。実際のパターンマッチングの際には、ノイズの影響から入力グラフやパターンに様々な変種が生じる。Bunke and Allermannはこの問題に対し、グラフの類似度を求めることで厳密性を除いたグラフマッチングを行っている。グラフの類似度を求める問題は、グラフの頂点と辺の置換・削除・挿入という基本的な編集操作を用いて二つのグラフを同型にする問題に帰着される。同型に変形するための一連の編集はコストが最小となるように行われ、そのときのコストがグラフの類似度とされる。

これらの手法はそれぞれ応用される特定の分野に強く依存しており、自然言語処理における依存構造の類似性評価への応用には適さない。Takahashi and Ishiyamaの手法では類似性評価の観点が構造的見地に大きく傾倒しており、各頂点(依存構造における単語)の類似性がグラフ類似性に与える影響を低く見積る傾向が見られる。事実、Takahashi and Ishiyamaの手法では、グラフを特徴づけるベクトル表現に頂点の情報がほとんど反映されない。各頂点が持つ情報としてベクトルに例外的に反映される情報は原子の質量数の総和などであり、自然言語の単語に関する情報をベクトルに反映することはほとんど不可能と言える。

また、Bunke and Allermannの手法では例えば図5.1のような構造を保持した部分木の移動に対して、辺の置換によるコストしか与えることができない。こうした性質は構文構造における選択制限の観点から、依存構造には適さない性質である。例えば図5.1における文“Fred saw a movie with Schwarzenegger.”[8]の依存構造において、前置詞句“with Schwarzenegger”に対応する構造の係り

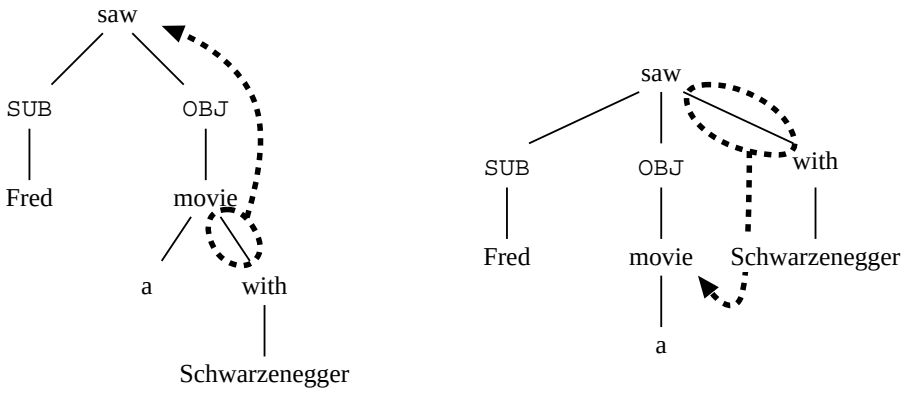


図 5.1: 前置詞句接続による解釈の変化

先が動詞 “saw” となるか名詞 “movie” となるかによって、文の意味は大きく異なる。Bunke and Allermann の手法による類似度には、こうした意味の違いを反映させることができない。

5.3 多義性解消の照合問題への応用

本研究で構築した多義性解消の手法は、5.2節に言及した先行研究が持つ問題を解消するものとして期待できる。これは主として次の理由によるものである。

- 3.3節に述べたように、本手法で用いる単語の配列は文の依存関係を良く保存するものである。
- 本手法は依存構造の類似性判断に、単語の配列を対象としたペアワイズアライメントを用いており、単語の類似性に基づいた類似性判断が可能である。
- ペアワイズアライメントは配列における要素の連接に高い評価値を与える。要素の連接は単語間の係り受け関係を意味するため、この性質は選択制限の観点から依存構造に適するものである。

こうした理由から、本手法の応用が文照合の精度を向上させることが期待できる。

本節では以上の考察に基づき、本研究で構築した多義性解消の手法を文照合の問題に応用することについて論じる。なお、本章で用いる単語の配列や

ペアワイズアライメント、およびそれに付随するいくつかの概念については、3.2節における定義に従うものとする。

今、任意の二つの文S、S'に対して構文解析を行った結果、それぞれの依存構造木として$D = (V, E)$、$D' = (V', E')$が導出されたものとする。このとき、D、D'に対応する単語の配列は定義3. 5 より一意に求められ、それぞれ$P = \{p_1, p_2, \cdots, p_m\}$、$P' = \{p'_1, p'_2, \cdots, p'_n\}$とする。このとき、二つの文$S$、$S'$の間の類似度を次の式で定義する。

$$simsent(S, S') = \frac{1}{2}(simdep(P, P') + simdep(P', P)) \tag{5.1}$$

ここで、$simdep(P, P')$は単語の配列の集合P、P'の類似度を表すものであり、次の式で定義される。

$$simdep(P, P') = \sum_{p_i \in P} \max_{p'_j \in P'} AS(p_i, p'_j) \tag{5.2}$$

$simdep(P, P')$は、3.4.2節における式(3.1)で定義した$sim(P_i, Q_{S_w})$に若干の変更を施したものになっている。$simdep(P, P')$では、P、P'は共に文の依存構造から得られた単語の配列である。すなわち、語義選択に対する配列パターンの貢献の度合いについては考慮する必要がなく、このために重みの付与を省いている。また、$simdep(P, P')$はPにおける特定の配列p_iに対して最も適合する配列をP'から求め、そのときの適合の度合いを$\max_{p'_j \in P'} AS(p_i, p'_j)$として算出し、これを$P$におけるすべての配列にわたって合計したものである。従って$simdep(P, P') \neq simdep(P', P)$の性質を持つ。この非可換の性質は式(3.1)における$sim(P_i, Q_{S_w})$に対しても成立するものである。式(3.1)では「多義語の文脈Q_{S_w}が語義選択の手がかりP_iにどれだけ類似するか」の観点で類似度を求めるため、非可換の性質は妥当なものとして作用する。しかし、文照合で用いる類似度は「文SとS'が相互にどれだけ類似するか」の観点で求められるべきである。このため、式(5.1)では$simdep(P, P')$と$simdep(P', P)$の平均を求めており、これをもって文の類似性を評価するものである。

5.4 実験

本節では、5.3節で述べた本手法の文照合への応用について、その妥当性を確認するために行った小規模な実験について述べる。本実験は文照合の応用として情報検索を想定し、検索質問(query)として与えられた文と類似する文

が正しく検索できるかどうかを検証するものである。この観点の下、本応用の妥当性検証として情報検索に関する次の二つの実験を設定した。

- 正解文をあらかじめ混入させたデータを検索対象とする文検索
- コーパス中に出現する文のうち、最も類似する文の組の抽出

以下、これらの小規模な実験について詳述し、それぞれの実験結果について言及する。

5.4.1 正解文の検索

本実験はあらかじめ用意した正解文を正しく検索可能かどうかを検証するものである。本実験ではまず検索質問の文とそれに対する正解文とをあらかじめ用意し、正解文をEDR Corpusから無作為に抜き出した1,000文の中に混入させて、1,001文からなる検索対象を作成する。これを対象に検索質問を検索した際に、正しく正解文を検索できるかどうかを調査した。ここで、EDR Corpusは出現する各文に形態素情報、構文情報などを付与した、いわゆるTreeBankの一種である。しかし本実験では、3.5節で言及した多義性解消の実験と同様に、EDR Corpusの各文に対してApple Pie Parserによる解析を施し、得られた結果に“SUB”、“OBJ”を自動的に付与しただけのもの(人手による修正を施さないもの)を各文の依存構造として利用する。これは、EDR Corpusで使用されている品詞体系への不満や、ノード“SUB”、“OBJ”の付与の問題など、いくつかの理由によるものである。

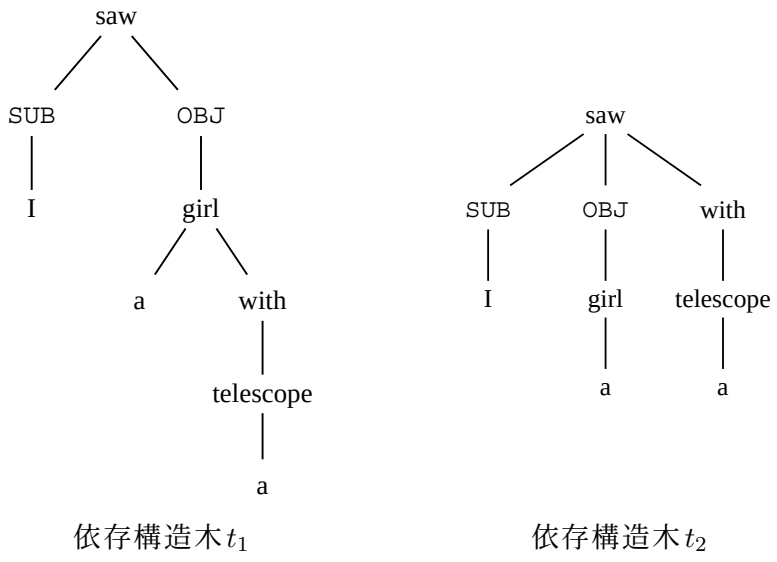
検索質問の文を“I saw a girl with a telescope”、正解文を“He has been found a boy with a hammer”としたときの検索結果を表5.1、5.2に示す。ここで、質問文には図5.2のように二通りの解釈t_1、t_2が存在するが、これらの表はそれぞれの依存構造木に対応した結果である。ここで注目すべき点は、表層文字列の基準では同一の質問文t_1とt_2が、依存構造を考慮することによって異なる検索結果を導いていることである。すなわち、正解文における前置詞句“with a hammer”の係り先は名詞“boy”であるため、類似した依存構造を持つ質問文t_1との間で高い類似度が得られている。このことは質問文の依存構造を考慮した検索によって、より精密な検索が可能となることを示唆するものである。

表 5.1: 質問文t_1による正解文の検索結果

順位	類似度	検索された文
1	9.914	He has been found a boy with a hammer.
2	6.290	I saw a people demonstrating for a ban on nuclear tests.
3	4.659	It looked a large book.

表 5.2: 質問文t_2による正解文の検索結果

順位	類似度	検索された文
1	6.290	I saw a people demonstrating for a ban on nuclear tests.
2	6.145	He has been found a boy with a hammer.
3	4.659	It looked a large book.



依存構造木t_1　　依存構造木t_2

図 5.2: 二通りの依存構造

5.4.2 類似文の組の抽出

本実験は同一コーパスの中から各文に最も類似した文を検索し、類似度が最も高い文の組を抽出するものである。本実験で用いるコーパスは、5.4.1節と同様にEDR Corpusから無作為に抜き出した1,000文で構成されるものとする。また、各文に対してApple Pie Parserによる解析を施し、得られた結果に“SUB”、“OBJ”を追加しただけのものを依存構造とする点も5.4.1節に同様である。

今、コーパスにおける各文を$S_1, S_2, \cdots, S_{1000}$で表すものとする。このとき本実験で文ごとに求める値Rは、導出の対象となっている文をS_iとすると$R_i = \max_{1 \leq j \leq 1000, j \neq i} simsent(S_i, S_j)$と表すことができる。本実験は$1 \leq i \leq 1000$について$R_i$を求め、このうち最も値の高い$R_i$と、$R_i$に対応する$(S_i, S_j)$の組を求めるものである。

図5.3に本実験で類似性を求めた組のうち、類似度の最も高い上位三つの組を示す。図5.3によると、これら三つの組はそれぞれ局所的に類似した構造を共有していることが観察される。直観的かつ定性的な評価ではあるが、コーパスの規模が小さいことなどを総合的に考慮に入れると得られた結果は極めて妥当なものであり、本応用の有効性を示唆しているものと評価できる。以下、得られた三つの組における類似性について述べる。

(a)の組は文“It also appears to have made a thorough check of Heiwa Sogo's nonperforming loans.”と“At face value, Eyres would appear to have a considerable head start on Skeggs.”の組である。この組では動詞句“appear to”を中心とした文の全体的な類似性が観察される。(b)の組は“Mercy killing is a big issue in the medical world.”と“This is a stark, white, squarish building in the central downtown area.”の組である。この組にも文全体を通じての類似性が観察される。また、単語“issue”と“building”はそれぞれ「出版」の意味と「建設」の意味を持ち、「営利活動」という観点で類似していると見なされることも(b)の高い類似度に影響している。(c)の組は“He was director and general manager of the International Finance Department of the parent Nomura.”と“Racloz, 44, has been president and general manager of Sandoz Venezuela since 1981”の組である。この組では“Someone is president/director and general manager of organization”という観点で極めて類似した文の組であり、こうした類似性が式(5.1)の類似度に反映されたものと考えられる。

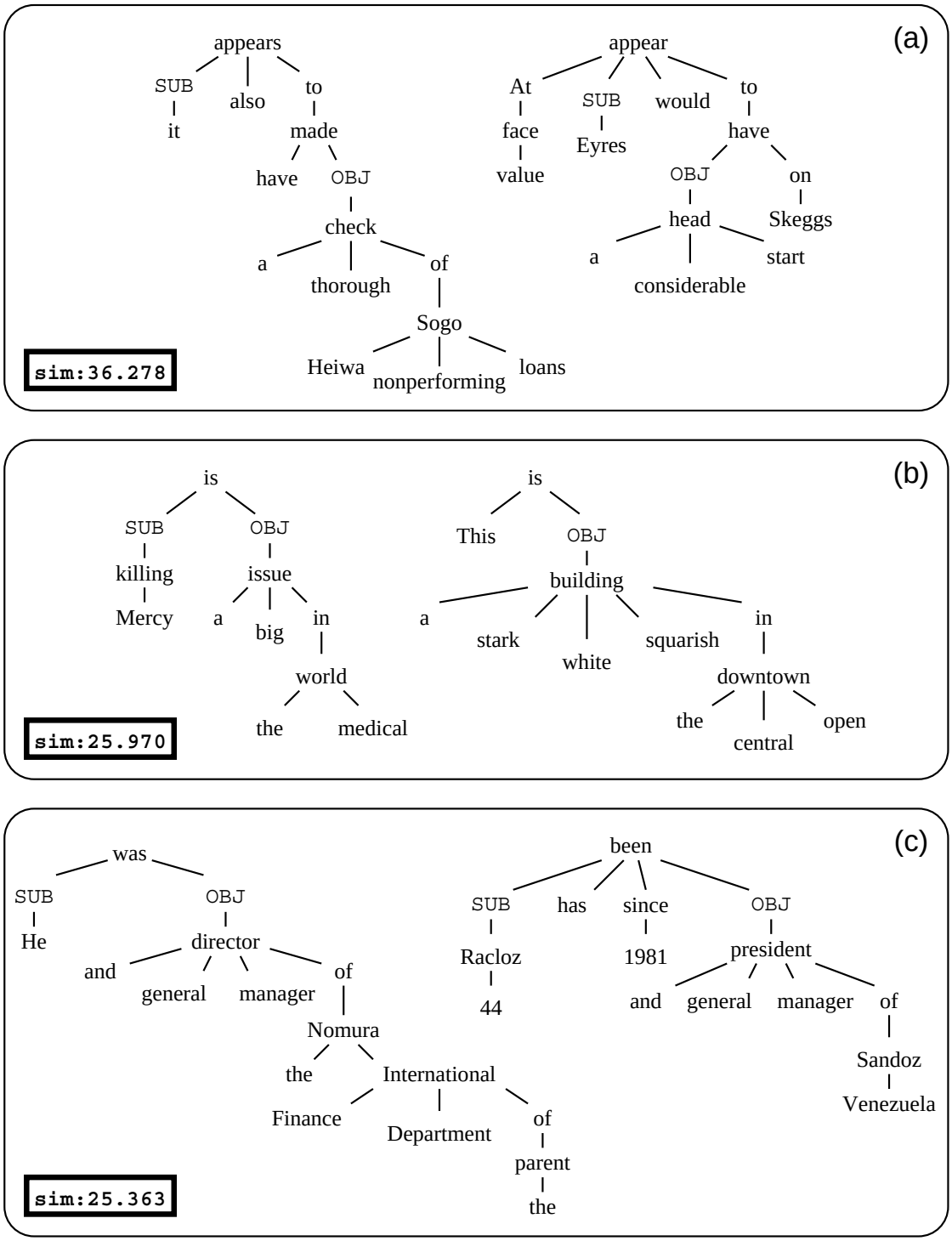


図 5.3: 最も類似する三つの文の組

5.5 まとめ

本章では多義性解消と情報検索の高い関連性に着目し、本研究で構築した多義性解消の手法を文照合の問題に応用することについて検討した。多義性解消の問題と情報検索の問題は一般に関連性が強いものと見なされ、両者に共通した技法が用いられる場合も多い。例えば多義性解消に関するSchützeの先行研究[48, 28]では文脈ベクトルの概念が導入されているが、これは情報検索における文書のベクトル表現の考え方を多義性解消に応用したものである。

本手法を文照合に応用することについては、次のような利点が考えられる。

- 文照合に表層文字列の情報だけでなく、依存構造の情報も用いることができる。
- 依存構造の表現として単語の配列を用いることにより、類似性に基づくグラフマッチングと比較して構造類似性評価の際の計算コストをある程度低くおさえることができる。
- 単語の配列を対象としたペアワイズアライメントを用い、WordNetを用いて配列要素の類似性を評価しているため、表層単語間の類似性を充分に考慮することができる。

こうした点から判断して、本手法の文照合への応用は有望なものと考えられる。

5.4節では、本応用の妥当性を検証するための二つの実験について述べた。両実験では妥当性の面で期待される結果が得られ、本応用が有望であることが示唆された。しかしこれらの実験は規模が小さく、実験結果の信頼性の面で充分とは言えない部分も存在する。また、実験結果の評価の定性的な側面も検討課題として挙げられる。自然言語処理を主な応用とする構造類似性評価の試み[49, 50, 51, 45]もいくつか報告されており、こうした報告における手法の評価方法を採用することで、本応用の定量的評価が可能になるものと考えられる。また、従来の手法と本応用との精度の面での比較によって、本応用の妥当性はさらに明らかになっていくものと期待される。

第 6 章

結論

本研究では自然言語が持つ曖昧性の一つである単語の多義性に対し、特に対象を動詞に絞り、これを解消するための新しい手法を構築した。多義性解消は、その代表的な応用として機械翻訳における訳語選択や情報検索における検索対象絞り込みなどが挙げられ、自然言語処理システムにおける有用性が極めて高い。このため、自然言語処理の最も初期の段階から多義性解消の問題は広く認識され、さまざまなアプローチによる研究が盛んに行われてきた。しかし、こうした研究の誕生から50年以上経った現在でも、単語の多義性の問題は充分に解決できているとは言えない。計算機とインターネットの急速な普及に伴なって、現在人間が相互に伝達している情報の量と多様性はこれまでにない速度で増大しつつある。こうした背景の下、計算機による自然言語情報の効率的な処理が強く求められており、高精度で高品質な自然言語処理システムが望まれている。本研究はこの要求に応えることを目標に行われたものである。本研究で構築された多義性解消の手法は、実装のコストの観点でいくつかの問題を有するものの、従来の手法よりも高い精度で多義性を解消することが可能である。本論文では、本研究で構築した多義性解消の手法の詳述、実装コスト軽減のための試みの報告、多義性解消以外の問題への応用の検討と、大きく三つについて論じた。本論文はこれらの論述を通じて、高精度・高品質の自然言語処理システム構築のための一手法を示すことができたものと考えている。

本論文では各章において具体的に次のことについて論述した。

第2章では、本論文で扱った多義性の問題の位置付けを明確にする目的で、自然言語処理における曖昧性の問題を概観した。曖昧性は主として自然言語の解析の際に問題となることに触れ、代表的な解析技術である形態素解析、

構文解析、意味解析について言及した。これらの記述を通じて、単語の多義性の問題が一般的に意味解析の一部として位置づけられることを明らかにした。また、本研究で扱った多義性解消の問題が具体的にどのような問題として定義されるかを明確にするため、語義の定義と多義の定義について言及し、単語の多義性の分類について説明した。ここではさらに、多義性解消の基本的概略についても説明した。次に、本研究で構築した手法の位置づけを明確にする目的で関連する先行研究について概説し、これらの先行研究が語義知識と学習方法の異なる二つの観点で分類できることを示した。こうした分類から、本研究で構築した手法が学習方法の観点では教師付き学習に基づく手法であることを示し、語義知識の観点からはこれまでの分類を混合した知識に基づく手法であることを説明した。

第3章では、本研究で構築した多義性解消の新しい手法について詳述し、本手法が従来の手法よりも高い精度で動詞の多義性を解消できることを示した。従来の手法は、語義知識の扱いの観点から連想関係に基づく手法と選択制限に基づく手法の二つに大別できる。これらの手法の欠点として、連想関係に基づく手法は選択制限を考慮することができず、選択制限に基づく手法は連想関係を考慮することができないという性質が挙げられる。ここではこの欠点の具体例として実際のコーパスに出現する文を挙げ、両手法が適切に多義性解消できない場合があることを明らかにした。本研究で構築した手法は語義知識として文の依存構造木全体から求められる単語の配列を用いるものであり、多義語に関する連想関係、選択制限を両者とも考慮できる。すなわち、本手法は連想関係を用いた手法の特長と選択制限を用いた手法の特長を併せ持った手法である。語義知識である単語の配列の類似性評価は、DNA配列の類似性評価に広く用いられているペアワイズアライメントの技法に基づいている。これによって、語義知識の類似度を柔軟かつ頑健に求めることが可能である。本章ではSENSEVAL-1の動詞を対象とした多義性解消に関する実験を説明し、実験結果から動詞全体の平均として81.1%の精度を達成したことを示した。個々の動詞の見地では、SENSEVAL-1への参加システムの最良の精度に比べて最大で14.2%の精度向上が見られた。こうした結果は本手法の妥当性が有意であることを示唆しており、動詞を対象とした場合に従来の手法よりも高い精度での多義性解消が可能であるという結論を導くことができる。

第4章では、本研究で構築した多義性解消の手法が有する問題の一つである高い実装コストを軽減する試みについて説明した。本研究で構築した多義性

解消の手法では語義知識の獲得が人手に依存しており、語義知識のコストが高いという欠点がある。人手によって獲得される知識には配列パターンと配列パターンに付与する重みとの二つがあるが、このうち配列パターンに付与する重みをコーパスからの統計情報に基づいて自動的に推定する試みについて述べた。本章では、訓練データにおけるパターンの適合と正解語義の選択の間の適合率、再現率から、閾値を帰納的に求める方法について説明した。また、得られた閾値を用いてパターンの適合・不適合を判断したとき、語義候補がどの程度絞り込まれるかという観点と、特定の語義候補の出現と特定パターンの出現は共起関係にあるかどうかという観点による二種類のエントロピーから、重みとペナルティを推定する方法について説明した。こうして得られた重みを用いた多義性解消の実験では、第3章と同じデータを用いることによって、SENSEVAL-1の動詞に対して平均71.3%の精度での多義性解消が達成できたことを示した。個々の動詞の見地では、SENSEVAL-1への参加システムの最良の精度に比べて最大で10.6%の精度向上が見られた。人手による重みを用いた場合には劣るものの、従来の手法よりも高い精度で多義性解消が可能であったという結果から、本章では本手法による重み推定が知識獲得のコストを削減する手法として有効であるという結論が導かれた。

第5章では、多義性解消と情報検索の間に高い関連性が存在することに着目し、本研究で構築した多義性解消の手法を文照合の問題に応用することについて検討した。本章ではまず文照合の問題について定義し、照合に表層文字列だけでなく文の構文情報まで用いることについて、用例に基づく機械翻訳や情報検索などの実際の応用の見地からその必要性を説明した。また、依存構造木の一般形であるグラフの類似性評価についての先行研究に触れ、これらの先行研究が自然言語処理への応用に適していないことについて言及した。次に、本研究で構築した多義性解消の手法を文照合の問題に応用した場合の変更点について検討し、本手法に基づいた文照合の定式化について述べた。さらに、本章における検討の妥当性を検証するために行った二つの小規模な実験について説明した。一つ目の実験についての記述を通じて、文照合に依存構造を用いることによって検索質問や検索対象の内容をより詳細に考慮できることと、より精密な情報検索が行えることを明らかにした。二つ目の実験についての記述からは、本応用による二つの文の間の類似度が直観的に妥当なものであることを示した。これらの実験結果から本応用が有望であることを示し、本応用のより大規模な評価実験を行うことの必要性を示すことがで

きた。本章を通じて、本研究で構築した手法が多義性解消以外の問題に応用することができることを示し、本手法の文照合に対する妥当性と、手法の一般性を明らかにした。

以上、本論文では本研究で構築した多義性解消の手法について論述し、さまざまな観点からの評価実験を通じて本手法の妥当性を明らかにした。後述するように本手法が持つ実装コストの問題は完全には解決できていないものの、自然言語処理システムに本手法を採用することによって多義性解消の精度を従来よりも向上させることが可能であり、従ってシステム全体の精度や品質を向上させることが期待できる。また、情報検索に本手法を応用することによってより精密な検索が実現可能となることが予想される。

本研究に残された課題のうち、最も大きな問題は実装コストの軽減である。第4章に述べた試みにより、少なくとも配列パターンに重みを与える部分のコストは削減することが可能となった。しかし、語義ごとに与える配列パターンを訓練データから自動的に構築するための手法はまだ確立できていない。配列上に見られるパターンを自動的に獲得する試みは分子生物学の分野で多数報告されている [38]。こうした試みが本研究に残された課題を解決するための糸口となることが期待される。

また、本研究の手法を動詞以外の品詞(名詞、形容詞、副詞)に適用することも積み残されている。本研究では問題の単純化のため対象を動詞に限定したが、本手法の一般性は文照合への応用においても明らかであり、特別な修正を施すことなしで動詞以外の品詞への一般化も可能であると予想される。こうした妥当性判断は大規模な実験によって明らかにする必要があるが、語義知識獲得のコストの問題から実験上の実装コストも高いため、本手法の大規模な評価実験は困難な傾向にある。語義知識の獲得が完全に自動化され、実装コストの軽減が実現できたときには、こうした大規模な評価実験が可能になるものと期待される。

謝辞

本研究のきっかけを与えてくださったのみならず、長い年月にわたり本研究に対して終始熱心にご指導いただきました元・静岡大学情報学部教授(現・浜松学院大学教授)・吉田敬一博士に厚く御礼申し上げます。また、本研究に対して懇切丁寧なご指導と適切なご助言をくださいました静岡大学情報学部教授・伊東幸宏博士に深く感謝申し上げます。本論文の作成にあたっては同大学同学部教授・梅谷征雄博士、同教授・北澤茂良博士、同助教授・小西達裕博士に数多くの有益なご指導をいただきました。心より感謝申し上げます。

また、論文誌「自然言語処理」の査読者の先生方からは本研究に対する有益なご指摘をいただきました。言語処理学会論文誌編集委員長・慶應義塾大学環境情報学部教授・石崎俊博士をはじめ、編集委員の諸先生方並びに査読者の先生方に心より感謝申し上げます。

本研究を進めるにあたっては、浜松大学学長・木宮一邦博士に研究活動に対するご理解とご支援をいただきました。浜松大学経営情報学部講師・片山清文博士からは研究活動を続けることに対してさまざまなご助言をいただきました。浜松大学電算機管理室・長崎洋康氏、同・西岡宏氏、同・望月祥正氏、常葉学園本部情報処理室・河合清和氏には電算機管理業務に多忙な中、数多くのご協力をいただきました。皆様に深く感謝申し上げます。

最後に、本研究の実施及び本論文の作成にあたり、筆者を精神的に支え色々と勇気づけてくれた妻・真由美に心より感謝いたします。

参考文献

[1] 情報処理学会（編）. 情報処理ハンドブック. オーム社, 東京都千代田区神田錦町3-1, 1997.

[2] Warren Weaver. *Translation*, pp. 15–23. Machine Translation of Language. John Wiley & Sons, New York, 1955. reprint of mimeographed version, 1949.

[3] 長尾確, 丸山宏. 自然言語処理における曖昧さとその解消. 情報処理, Vol. 33, No. 7, pp. 746–756, 1992.

[4] 麻野間直樹, 中岩浩巳. 目的言語の単語共起情報を利用した訳語選択と未知語の訳出. 言語処理学会第5回年次大会発表論文集, pp. 442–445. 言語処理学会, March 1999.

[5] Hinrich Schütze and Jan O. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175, Las Vegas, 1995.

[6] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 4.0. 東京大学大学院情報理工学系研究科, July 2003.

[7] 長尾真. 自然言語処理. 岩波書店, 東京都千代田区一ツ橋2-5-5, 1996.

[8] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, Massachusetts 02142, 1993.

[9] 島津明, 内藤 昭三浩郷. 助詞「の」が結ぶ名詞の意味関係の解析. 計量国語学, Vol. 15, No. 7, pp. 247–266, 1986.

[10] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts 02142, 1999.

[11] Mark Stevenson. *Word Sense Disambiguation: The Case for Combination of Knowledge Source*. CSLI Publications, Stanford, California, 2003.

[12] Stephen F. Weiss. Learning to disambiguate. *Information Storage and Retrieval*, Vol. 9, pp. 33–41, 1973.

[13] Abraham Kaplan. An experimental study of ambiguity and context. *Mechanical Translation*, Vol. 2, No. 2, pp. 39–46, 1955. reprint of mimeographed version, 1950.

[14] William A. Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, Vol. 26, pp. 415–439, 1993.

[15] Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of art. *Computational Linguistics*, Vol. 24, No. 1, pp. 1–40, 1998.

[16] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, Vol. 3, No. 4, pp. 235–244, 1990.

[17] David Yarowsky. Word sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 454–460. COLING, 1992.

[18] Longman dictionary of contemporary english, third edition, January 1978.

[19] Jean Véronis and Nancy Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, pp. 389–394, 1990.

[20] Yoshiki Niwa and Yoshihiro Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 304–309, 1994.

[21] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word sense disambiguation using statistical methods. In

Proceedings of the 29th Annual Meeting of Association for Computational Linguistics, pp. 264–270, 1991.

[22] Marti A. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Center for the New OED and Text Research*, pp. 1–22, 1991.

[23] Dekang Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics*, pp. 64–71, 1997.

[24] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings 29th Annual Meeting of the Association for Computational Linguistics*, pp. 169–176, Berkeley, June 1991. Association for Computational Linguistics.

[25] 後藤斉. 言語理論と言語資料 – コーパスとコーパス以外のデータ, 日本語学, 第22巻, pp. 6–15. 明治書院, 4月臨時増刊号「コーパス言語学」, 2003.

[26] 日本電子化辞書研究所. EDR電子化辞書 1.5版 使用説明書, 1996.

[27] Jiri Stetina and Makoto Nagao. General word sense disambiguation method based on a full sentential context. 自然言語処理, Vol. 2, No. 5, pp. 47–74, 1998.

[28] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, Vol. 24, No. 1, pp. 97–123, 1998.

[29] 福本文代, 辻井潤一. コーパスに基づく動詞の多義解消. 自然言語処理, Vol. 4, No. 2, pp. 21–40, 1997.

[30] David Yarosky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, Cambridge, 1995. Association for Computational Linguistics.

[31] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discource. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 233–237, 1992.

[32] David Yarowsky. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, Vol. 34, pp. 179–186, 2000.

[33] Adam Kilgarriff. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Vol. 1, pp. 581–585. LREC, 1998.

[34] Adam Kilgarriff and Joseph Rosenzweig. English senseval: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Vol. 3, pp. 1239–1244, Athens, Greece, 2000. LREC.

[35] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[36] 後藤修. 核酸・蛋白質一次構造の計算機による解析. 日本物理学会誌, Vol. 38, No. 6, pp. 477–480, 1983.

[37] Christiane Fellbaum. English verbs as a semantic net. *International Journal of Lexicography*, Vol. 3, No. 4, pp. 270–301, 1990.

[38] 美宅成樹, 金久實. ヒトゲノム計画と知識情報処理. 培風館, 東京都千代田区九段南4-3-12, 1995.

[39] Satoshi Sekine. *Corpus-Based Parsing and Sublanguage Studies*. PhD thesis, New York University, 1998.

[40] I. Dan Melamed and Philip Resnik. Tagger evaluation given hierarchical tag sets. *Computers and the Humanities*, Vol. 34, pp. 79–84, 2000.

[41] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 268–275, 1991.

[42] Philip Resnik. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-based Natural Lanugage Processing Techniques*, pp. 56–64, 1992.

[43] Philip Resnik. Using information content to evaluate sementic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, Montreal, August 1995. IJCAI.

[44] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 29–34, Pittsburgh, 2001. Association for Computational Linguistics.

[45] 高橋哲朗, 乾健太郎, 松本裕治. テキストの構文的類似度の評価方法について. 情報処理学会研究報告, No. 150-NL-24. 情報処理学会, 2002.

[46] Yoshimasa Takahashi and Yuichi Ishiyama. *Evaluation of Molecular Similarity Using Topological Fragment Spectra*, pp. 311–316. Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches. Kluwer Academic Publishers, Netherlands, 1995.

[47] H. Bunke and G. Allermann. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, Vol. 1, No. 4, pp. 245–253, 1983.

[48] Hinrich Schütze. Word sense disambiguation with sublexical representations. In C. Weir, S. Abney, R. Grishman, and R. Weischedel, editors, *Workshop Notes, Statistically Based NLP Techniques*, Vol. 1, pp. 109–113. AAAI, 1992.

[49] David T. Barnard, Gwen Clarke, and Nicholas Duncan. Tree-to-tree correction for document trees. Technical Report 95-372, Department of Computing and Information Science, Queen's University, Kingston, Ontario K7L 3N6, 1995.

[50] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems (NIPS'2001)*, pp. 625–632, 2001.

[51] Michael Collins and Nigel Duffy. Parsing with a single neuron: Convolution kernels for natural language problems. Technical Report UCSC-CRL-01-01, University of California at Santa Cruz, 2001.

[52] A. Nádas, D. Nahamoo, M. A. Picheny, and J. Powell. An iterative “flip-flop” approximation of the most informative split in the construction of decision trees. In *Proceedings of the 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP91)*, Vol. 1, pp. 565–568, the Sheraton Centre Hotel & Towers, Toronto, Ontario, Canada, May 1991. Acoustics, Speech and Signal Processing Society, the Institute of Electrical and Electronics Engineers.

[53] 辻井潤一. 自然言語理解の歴史と現状. 情報処理, Vol. 30, No. 10, pp. 1142–1149, 1989.

[54] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, 1990.

[55] Dekang Lin. Word sense disambiguation with a similarity-smoothed case library. *Computers and the Humanities*, Vol. 34, pp. 147–152, 2000.

[56] Adam Kilgarriff and Joseph Rosenzweig. Framework and results for english senseval. *Computers and Humanities*, Vol. 34, pp. 15–48, 2000.

[57] 北本朝展, 高木幹雄. 類似画像検索への応用を目的とした階層化構造付きグラフマッチングの高速化. 画像の認識・理解シンポジウム (MIRU'96), 第II巻, pp. 331–336, 1996.

[58] Victor H. Yngve. *Syntax and the Problem of Multiple Meaning*, pp. 208–226. Machine Translation of Language. John Wiley & Sons, New York, 1955.

[59] Martin Chodorow, Claudia Leacock, and George A. Miller. A topical/local classifier for word sense identification. *Computers and the Humanities*, Vol. 34, pp. 115–120, 2000.

[60] Koichi Yamashita, Keiichi Yoshida, and Yukihiro Itoh. Word sense disambiguation using pairwise alignment. In *ACL-03 Companion Volume to the Proceedings of the Conference*, pp. 157–160, 2003.

[61] 白井清昭. Senseval-2 日本語辞書タスク. 自然言語処理, Vol. 10, No. 3, pp. 3–24, 2003.

付録 A

依存構造木へのノード “SUB”、“OBJ” の追加

本研究では任意の依存構造木(V, E)に対してノード “SUB”、“OBJ” を追加している。このノードは動詞の主格と目的格の違いを明確にするために追加しており、“SUB” は主格、“OBJ” は目的格を表している。ノードの挿入の手順は図A.1に示す。ここで、$V = \{w_1, w_2, \cdots, w_N\}$とする。

begin
  for each verb $w_v \in V$ **do begin**
    for each noun w_n that $w_n \in V$ and $(w_v, w_n) \in E$ **do begin**
      delete (w_v, w_n) from E;
      if $(n > v)$ **then begin**
        add "OBJ" to V;
        add $(w_v$, "OBJ") and ("OBJ", $w_n)$ to E; **end**;
      else if ((w_v is past participle) and
            (there is be-verb w_b that $w_b \in V$ and $(w_v, w_b) \in E$)) **then begin**
        add "OBJ" to V;
        add $(w_v$, "OBJ") and ("OBJ", $w_n)$ to E; **end**;
      else begin
        add "SUB" to V;
        add $(w_v$, "SUB") and ("SUB", $w_n)$ to E; **end**;
    end;
  end;
end.

図 A.1: 依存構造へのノード "SUB"、"OBJ" の追加手順

付録 B

ペアワイズアライメントの導出

最適なアライメントを求める手法には、動的計画法に基づく手法、有限状態オートマトンに基づく手法、隠れマルコフモデルに基づく手法などさまざまな手法がある。ここでは動的計画法に基づく手法を示す。任意の二つの配列を$p = (w_{p,1}, w_{p,2}, \cdots, w_{p,m})$、$q = (w_{q,1}, w_{q,2}, \cdots, w_{q,n})$とする。$p$、$q$間の最適なペアワイズアライメントを求めるアルゴリズムを図B.1に示す。また、このアルゴリズムを用いて配列(`cake,made,at,home`)と(`meeting,at,room,102`)とのペアワイズアライメントを求める様子を図B.2に示す。

本論文におけるペアワイズアライメントは、単語の配列の長さが一様でないことを考慮し、アライメントの左右両端にあるギャップにペナルティを与えないよう、アルゴリズムに以下の変更を加えている。まず、図B.1のアルゴリズムにおいて、初期値の付与を次の式に従うよう変更する。

$$F_{i,0} = 0 \quad (0 \leq i \leq m) \tag{B.1}$$

$$F_{0,j} = 0 \quad (0 \leq j \leq n) \tag{B.2}$$

また、pとqのアライメントスコア$AS(p,q)$は次のように定義する。

$$AS(p,q) = \max\{F_{i,n}, F_{m,j} | 0 \leq i \leq m, 0 \leq j \leq n\} \tag{B.3}$$

式 (B.3) で選択されたノードをF_{max}とすると、最適なアライメントはmaxに対応するノードからマトリクスの遷移を逆向きに辿ることによって求められる。

begin

prepare $(m+1) \times (n+1)$ matrix F, R, X, Y;

/* 初期値の付与 */

$F_{0,0} = 0$; $R_{0,0}$ =nil;

for $i = 1$ **to** m **do begin** $F_{i,0} = d(w_{p,i}, \text{“-”}) * i$; $R_{i,0} = horizontal$; **end**;

for $j = 1$ **to** n **do begin** $F_{0,j} = d(\text{“-”}, w_{q,j}) * j$; $R_{0,j} = vertical$; **end**;

/* DPマトリクスの各ノードの計算 */

for $i = 1$ **to** m **do for** $j = 1$ **to** n **do begin**

$F_{i,j} = \max\{(F_{i-1,j} + d(w_{p,i}, \text{“-”})), (F_{i,j-1} + d(\text{“-”}, w_{q,j})), (F_{i-1,j-1} + d(w_{p,i}, w_{q,j}))\}$;

if (selected node is $F_{i-1,j}$) **then begin**

$R_{i,j} = horizontal$; $X_{i,j} = w_{p,i}$; $Y_{i,j} = \text{“-”}$; **end**;

else if (selected node is $F_{i,j-1}$) **then begin**

$R_{i,j} = vertical$; $X_{i,j} = \text{“-”}$; $Y_{i,j} = w_{q,j}$; **end**;

else if (selected node is $F_{i-1,j-1}$) **then begin**

$R_{i,j} = oblique$; $X_{i,j} = w_{p,i}$; $Y_{i,j} = w_{q,j}$; **end**;

end;

$AS(p,q) = \max\{F_{i,n}, F_{m,j} | 0 \leq i \leq m, 0 \leq j \leq n\}$; /* アライメントスコアの獲得 */

/* 最良の経路の探索 */

set i, j to point to selected node F_{max};

$p' = \phi$; $q' = \phi$;

while $(R_{i,j} \neq nil)$ **begin**

add $X_{i,j}$ to p' as the first element;

add $Y_{i,j}$ to q' as the first element;

if $(R_{i,j} = horizontal)$ **then** $i = i - 1$;

else if $(R_{i,j} = vertical)$ **then** $j = j - 1$;

else if $(R_{i,j} = oblique)$ **then begin** $i = i - 1$; $j = j - 1$; **end**;

end;

end.

図 B.1: 動的計画法に基づいたペアワイズアライメント

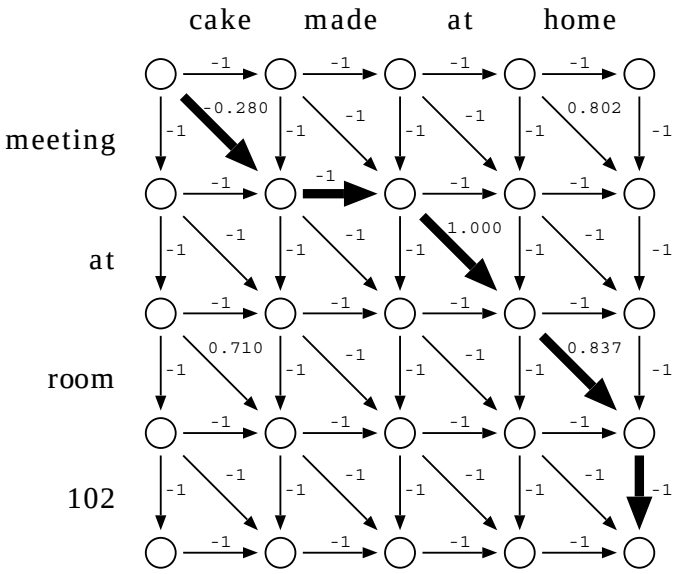


図 B.2: ペアワイズアライメントの例