論文の内容の要旨

論文題目　**Automatic Construction of Word Sense Association Networks**

（語義関連ネットワークの自動生成）

氏名　梶　博行

本論文は，自然言語の意味処理の基盤となる「語義関連ネットワーク」を対象分野のテキストデータの集合体であるコーパスから自動生成する枠組みを提案し，そのために必要な技術を論じたもので，全６章から構成されている．

第１章「緒言」では，研究の背景と目的を述べるとともに関連分野における従来研究を概観する．自然言語処理システムでは，語の文法的・意味的な知識を記述した辞書が重要な役割を果たす．しかし，膨大な語彙と言語現象の多様性・個別性のため，辞書の作成が実用的なシステムを実現する上での隘路となっている．他方，電子化されたテキストデータが大量に利用できるようになり，コーパスに基づく自然言語処理やコーパスからの知識獲得の研究が進展している．このような状況に鑑みて，本論文では，高精度の自然言語処理に必要な意味辞書を対象分野のコーパスから自動生成する方法を提案する．２言語の言語リソースを用いることが提案方法の特徴である．概念から語への写像が言語によって異なることを利用して，２言語コーパスと対訳辞書から概念（語義）自体を自動抽出する．本方法では第２の知識源である対訳辞書の網羅性が重要であるので，２言語コーパスからの対訳語抽出に関しても新しい方法を提案する．本研究では，また，適用分野の広さという観点か

ら，対訳関係をもつパラレルコーパスではなく，緩やかなコンパラブルコーパスすなわち分野が同じという以外に特別な関係のない２言語のコーパスの組に適用可能な技術を開発することを目標とする．本章の後半では，本研究に関連する「意味辞書」，「コーパスからの対訳語の抽出」，「コーパスに基づく語義の曖昧性解消」，および「コーパス中の分布に基づく語のクラスタリング」に関する技術を概観する．

第２章「語関連から語義関連へ：２言語コーパスを用いる１アプローチ」では，語義関連ネットワーク自動生成の全体的な枠組みについて述べる．語義関連ネットワークは，節点が２言語の同義語集合で定義される語義を表し，枝が語義間のトピック的な連想関係を表すグラフである．語義関連ネットワーク自動生成の基本アイデアは次のとおりである―コンパラブルコーパスを構成する各言語のコーパスからそれぞれの言語の語関連ネットワークを生成し，対訳辞書を介してそれら２つの語関連ネットワークを対応付ける（align）ことによって，語を語義に分割するとともに語の関連を語義の関連に変換する．語の語義への分割と語関連の語義関連への変換は不可分であるが，一度に行なうことは困難である．そこで，語義関連ネットワークの中間形式として，個々の語に対する語義リストと語義 - 手がかり語相関行列を設定する．語義リストは，訳語集合によって定義された語義のリストである．語義 - 手がかり語相関行列は，語義と手がかり語（語義を同定する手がかりになるので，関連語を手がかり語（clue）と呼ぶ）の相関を表す行列である．この中間形式データを用いて，与えられた語義リストに対し語義 - 手がかり語相関行列を計算する「語義 - 手がかり語の相関計算」（第４章），語義 - 手がかり語相関行列を利用して語義リストを生成する「訳語集合のクラスタリング」（第５章）を交互に実行する．このようにして個々の語の語義リストと語義 - 手がかり語相関行列を求めたあと，すべての語の中間形式データを統合することによって語義関連ネットワークを生成する．本章では，機械翻訳，情報検索を始めとする様々な自然言語処理システムにおける語義関連ネットワークの必要性，また，相補的な関係にある概念分類型の意味辞書 WordNet との結合可能性にも言及する．

第３章「文脈類似度に基づく対訳語の抽出」では，２言語コーパスから対訳語ペアを抽出する新しい方法を提案する．対訳語抽出の従来技術として，パラレルコーパスの対応する文中に共に出現する頻度などを利用する統計的方法，単純語の対訳辞書を用いて複合語の構成要素間の対応関係をチェックする言語的方法があるが，それぞれ一長一短がある．本研究では，両言語の語を共起頻度付きの共起語集合で特徴付け，基本語の対訳辞書を用いて共起語集合の類似度を計算し，類似度の高い

語のペアを抽出する方法を開発した．この方法は，コーパスが与える共起データと対訳辞書が与える対訳知識をうまく結び付けることにより，統計的方法と言語的方法の長所を併せ持つ方法となっている．すなわち，単純語ペア，構成要素レベルの対応関係が成立しない複合語ペアを含む様々なタイプの対訳語ペアを抽出することができる．また，文の対応付けが困難な２言語コーパスに適用可能であり，小さなコーパスから低頻度の対訳語ペアも抽出することができる．本方法を評価するため，日本語と英語の対応する特許明細書から対訳辞書に含まれていない対訳語ペアを抽出する実験を行なった．その結果，頻度１以上の対訳語ペアに対して再現率が 33.8%，精度が 76.7%であり，対訳辞書のカバー率を高める方法として有効であることを確認した．

第４章「語関連の言語間対応付けに基づく語義 - 手がかり語相関の反復計算」では，多義語の語義と語義を同定する手がかり語との相関をコンパラブルコーパスに基づいて計算する方法を提案する．語義と手がかり語の相関を求めることは，語義の曖昧性解消手法の学習フェーズと考えることができる．コンパラブルコーパスを用いる語義の曖昧性解消の研究は先例がないが，近い例として第２言語の単言語コーパスと対訳辞書を用いる語義の曖昧性解消がある．しかし，その方法は学習フェーズをもたず，また，言語間でのカバーするトピックのずれやデータスパースネスの問題が考慮されていない．本研究では，これらの問題を克服するため，２つの仮説―語関連の対応付けの確度は当該語関連に随伴する関連語の間の対応付けの確度に依存する．また，語義と手がかり語の相関は当該語義と随伴する手がかり語との相関に依存する―に基づいて，語義と手がかり語の相関を反復計算するアルゴリズムを考案した．この反復計算はスパースなデータに対するスムージング効果をもち，コーパス中のインスタンスの数が比較的小さな多義語にも適用可能な方法となっている．全インスタンスの文脈データが圧縮された語関連の集合から学習するので収束が速く，計算量の面でも十分実用的である．本方法を評価するため，計算結果である語義 - 手がかり語相関行列を用いた語義の曖昧性解消実験を行なった．すなわち，テキスト中の多義語に対して，文脈中の手がかり語との相関に当該多義語からの距離に応じた重みを掛けた値の和として定義されるスコアが最大となる語義を選択した．学習に使用したコーパスはウォールストリートジャーナル 1.5 年分と日本経済新聞 1 年分である．対訳辞書は EDR（日本電子化辞書研究所）対訳辞書を用いた．ベースライン（コーパス中の最頻語義を文脈に関係なく選択する方法）の精度 62.8%に対して，適用率 95.8%，精度 76.2%という結果であり，本方法の有効性を実証する

ことができた．

第５章「原言語での分布パターンの類似度に基づく訳語のクラスタリング」では，コンパラブルコーパスに基づいて多義語の語義を獲得する方法を提案する．語義の自動獲得は，恣意性の排除，分野に依存した語義の定義といった効果が期待される重要な課題である．先行研究は少ないが，単言語コーパス中の分布パターンに基づく語のクラスタリングによって，同義語の集合で定義される語義を獲得する方法が提案されている．これに対して，本研究では，第４章の語義 - 手がかり語相関計算方法を用いて原言語に写像した分布パターンの類似度に基づいて，多義語に対する訳語の集合を階層的にクラスタリングする方法を開発した．類似度を計算する相手の訳語あるいは訳語集合を除外して計算した語義 - 手がかり語相関行列を併用することによって，訳語の使用頻度の差の影響を小さくする工夫を加えた．本方法によれば，訳語が多義語であっても，目的とする多義語の訳語としての分布パターンで特徴付けられる．また，語義 - 手がかり語相関の反復計算の結果，使用頻度の低い訳語も密な分布パターンで特徴付けられる．このため，コーパス中で使用されている語義のみを精度よく獲得することができる．本方法は，原言語での比較的小さな次元の分布パターンで特徴付けられた少数の訳語をクラスタリングするので，計算量の面でも実用的である．第４章と同じコンパラブルコーパスと対訳辞書を用いた評価実験の結果は，コーパス中の使用比率が 5%以上の語義に関する語義再現率が 87.1%，クラスタリング結果の適切さを示す指標として提案した語義定義の精度が 76.6%であり，本方法のフィージビリティを確認することができた．

第６章「結言」では，研究の成果をまとめ，今後の方向について述べる．本研究では，意味処理の基盤となる語義関連ネットワークを２言語コンパラブルコーパスから自動生成するための基本技術を開発した．今後の方向として，より高精度の語義関連ネットワークを生成するため，構文解析技術の利用を検討することが重要である．また，語義関連ネットワークを様々な自然言語処理応用システムに適用してその有効性を実証する．

以上のように，本研究では，これからの自然言語処理において重要な意味辞書を対象分野のテキストコーパスから自動生成する手法を提案し，そのフィージビリティを実証した．緩やかなコンパラブルコーパスからの知識獲得という困難な課題に対して道を拓いており，自然言語処理技術の発展に少なからず寄与するものと考える．