報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Modeling and Selection of Context for Better Synonym Acquisition
(類義語獲得の性能向上のための文脈のモデル化と選択)

氏　名　萩原 正人

論　文　内　容　の　要　旨

語彙に関する知識は，自然言語処理において基本的かつ重要な言語資源の一つであり，情報検索におけるクエリ拡張や，シソーラスの自動構築など，幅広い応用がある．しかし，人手により語彙知識体系を構築・維持するコストは高いため，計算機によるサポートや自動化が必要不可欠である．

様々な語彙関係の中でも，類義語関係は様々なアプリケーションにおいて用いられる基本的な関係である．語の類似関係をコーパスから自動的に検出するために，分布類似度と呼ばれる概念がこれまで広く用いられてきた．分布類似度は，「意味の類似している語は類似した文脈を有する」という「分布仮定」に基づき，コーパスから抽出された語の文脈の共通度を用いて計算される類似度である．

しかしながら，これまでの研究の多くが類似度の計算方法に注目したものであり，分布類似度の本質である文脈自体の重要性については，ほとんど注目されてこなかった．具体的には，類義語獲得のために有効な文脈の定式化および比較，ましてや拡張については，ほとんど議論されていないという問題がある．しかし，分布類似度を用いる際に，どのように文脈を構築・拡張するかは基礎的かつ重要な問題であると考えられる．また，類義語獲得にとって有効な文脈はこれまで経験的にのみ決められてきたが，これは類義語獲得の性能および計算量に大きく影響する重要な要因である．さらに，従来，単純なベクトル空間モデルが主に用いられてきたが，より高度な意味表現手法が類義語獲得にとって有効であると考えられる．

従来研究において注目されてこなかったこれらの問題に対し，本論文では，より高性能な類義語獲得を目指して，文脈の定式化，拡張，選択，モデリングについて改善手法を提案し，それらが類義語獲得において有効であることを示すことを目的とする．具体的には，本論文の貢献は以下の３点である．まず第一に，文脈の表現および拡張の問題に対し，従来の依存関係に基づく文脈を定式化し，間接依存関係

に基づいた拡張が有効であることを示す．第二に，文脈の選択の問題に対し，コーパスから抽出された文脈を，統計的指標に基づいて取捨選択するための３つの選択手法を提案し，性能/コストのトレード・オフの観点から有効性を比較・検討する．第三に，文脈モデリングの問題に対し，語と文脈の共起のモデル化のために潜在意味モデルを適用し，これらのモデルが類義語獲得の性能を向上させることを示す．また，距離学習および文脈素性に基づいた分類という２つの教師有り学習手法を提案し，従来の教師無しアプローチと比較してより高い精度で類義語獲得が可能になることを示す．なお，本論文では英語を対象としているが，ここで提案する手法は特定の言語に限定されない．

本論文は全８章から構成される．第１章は本論文の序章であり，分布類似度に関連した諸問題について論じ，本論文の立場を明らかにする．

第２章ではまず，分布類似度に関する基本的概念を紹介する．続いて，文脈情報である依存関係をコーパスから抽出する手法について説明し，従来提案されてきた類似度指標と重み関数をいくつか提示する．これらの要素技術は，本論文における実験のベースラインとして用いるものである．

第３章では，第２章において提示した概念を，類義語獲得のタスクに実際に適用する．類義語獲得の性能を評価するために，WordNet 等既存のシソーラスを参照として用いる２つの評価指標，平均精度 (AP) および相関係数 (CC) を提案する．まずはじめに，実験の前処理として，語および文脈の出現頻度による足切りの効果を調査した．続いて，類似度指標および重み関数の性能を比較し，文脈ベクトルや確率分布に対して何らかの正規化を施した類似度指標であるコサイン類似度，ベクトル間 Jaccard 係数，Jensen-Shannon 距離の性能が比較的良いことを示した．また，重み関数の中では PMI や t-test の性能が最も良いことがわかったが，これは，語単位での正規化係数が貢献していると考えられる．

第４章では，２つの潜在意味モデル LSI および PLSI を類義語獲得へと適用し，単純なベクトル空間モデルと比較する．潜在意味モデルは，語と文脈の共起を潜在意味を介してモデル化することにより，スパースネスやノイズの問題に対処することができる．性能の振る舞いはモデルごとに大きく異なっていたものの，適切な類似度指標と組み合わせることにより従来手法よりも高い獲得性能を示した．また，性能は潜在意味クラス数が約 100 から 150 でピークに達することが分かった．

第５章では，分布類似度に通常用いられる直接的な依存関係を定式化したのち拡張し，間接的にしか関連していない語も含められるよう，さらに高度な文脈情報である間接依存関係の利用を提案する．実験により，直接依存関係に加えて間接依存関係も利用することにより類義語獲得の性能が向上することを確認した．また，間接依存関係を用いる際の文脈表現についても比較・検討し，性能の向上は，粒度の細かい文脈表現を用いた際に顕著であることを示した．

第６章では，カテゴリ，文脈タイプ，共起のそれぞれに基づいた３種類の文脈選択手法を提案し，実験により各選択手法および各指標の特徴を明らかにした．結果として，最も単純であるカテゴリに基づいた選択手法でも，十分効果的であることが分かった．一方，文脈タイプおよび共起に基づいた選択手法は，近接関係・依存関係どちらの文脈情報に対しても有効であり，いかなる文脈や次元数/計算量の制約下においても汎用的に用いることができることを示した．

第７章では，教師有り学習に基づく類義語獲得手法を２つ，新たに提案する．一つ目は分類モデルであり，「分布素性」と呼ぶ文脈に基づく素性を提案し，それに加えて，語のペアに対する文脈パターンに基づく素性の両方を用いることにより，完全に統合された分類器を構築することができる．評価実験の結果，分布類似度と比較して，分布素性によってＦ値ベースで性能を60%以上向上させることができることを示した．二つ目のアプローチは，ユークリッド距離の一般化であるマハラノビス距離を訓練データから学習するものであり，既存の類似度指標と比較して有意に良い性能を示した．依然として，素性選択により次元数を大幅に削減する必要があるが，教師有り学習により得られる性能の向上はこの欠点を補い，いくつかのアプリケーションにおける使用を正当化するのに十分であると考えられる．

第８章は本論文の結びであり，各章において得られた結果と知見を総括し，将来の展望について示す．