博士論文要旨

２００８年　1月　4日

申請者

氏　名　中渡瀬（なかわたせ）　秀一（ひでかず）

論文題目　　テキストコーパスを用いた語の相互関係の発見に関する研究

要旨

本論文は、日本語コーパスを解析し、語の相互関係を発見するための手法を提案し、その有効性の検証について述べる。近年、電子化文書が多く蓄積され、ネットワークを通じてそれらを検索、分類、加工する利用が進んでいる。その際には、それらの文書は自然言語処理によって解析が行われる。そしてこの解析を行う時には語彙知識の電子化辞書が必要とされる。語の相互関係はこの語彙知識の一種であり、電子化辞書の重要な構成要素である。本論文では、そのような語の相互関係として類義関係、多義関係、上位下位関係を対象にし、それらを日本語コーパスの解析に基づいて発見する手法の提案とその検証を行った。

コーパスから本研究で扱う語の相互関係を発見するときに、以下の２つの課題が存在していた。第一に、ある2語の間に類義関係が存在するか否かは観点に依存しているため、どの観点による類義関係なのかをいかに明確にするかというものである。従来の類義語獲得の研究では、主に語の間の類似度に注目したクラスタリングによる手法が用いられてきた。しかしこれらの研究は、2語の間には常に一意な類似度が与えられるため類似性の観点による違いを扱うことが困難であった。そして類義語集合の作成はこの類似度に基づくため、ある語A、B、Cに関してAB間の類似度とAC間の類似度の観点が同じでないときにも同一尺度として、それを基準としてクラスタリングするために不適切な類義語集合が得られるという問題があった。

第二に、語の間の上位下位関係を発見するための手法について、その導出をどのように上位下位関係の定義から合理的に行うかというものである。この点で従来の上位下位関係獲得の研究では、コーパス中に存在する「AなどのB」といった定型表現からBとAの上位下位関係を抽出する手法などが提案されていた。この場合、抽出対象となる定型表現にはその著者の意識した上位下位関係しか表現されないので、著者が意識しないもしくは、自明であるが電子化辞書に必要な上位下位関係が得られないという問題があった。それに対して、コーパス中の語の共起関係に基づく抽出手法も提案されている。例えば語A、Bが類義関係にあり、Aの出現頻度がBより高ければAはBの上位語とする手法のように、上位下位関係の定義から導出されないために、それらは合理性を欠いていた。

第一の課題に対しては、同じ観点で類義関係にある語だけを直接グルーピングして類義語集合を得るというアプローチを検討し、その新たな手法として、語の共起関係から得られるグラフ構造から類義語集合なすグラフ構造を探索し列挙する方法を提案し、検証実験によって、提案手法が同じ観点の類義関係をグルーピングすることにおいて有効であることを示した。

(備考)　１．和文で作成する場合は 2,000 字～3,000 字，英文で作る場合は 700 語～2,000 語程度で作成すること。

２．用紙の大きさは，日本工業規格 A4 縦型とすること。

第二の課題に対しては、まず名詞が指し示す対象の範囲の包含関係によって語のIS-A関係を定めることを示した。次にある動詞に対して格関係をもつ対象も名詞として表現されることに着目し、動詞と名詞の依存関係を用いてIS-A関係を発見する方法を検討した。そして動詞と名詞の依存性解析を行って得られる名詞ごとの動詞集合を比較する手法の提案を行った。さらに、この手法の有効性を検証するための実験を行い、本手法によって獲得されたIS-A関係が従来手法によるカバレッジを大幅に拡大することが確認した。

本論文は６章からから構成される。

第１章では、本研究における背景と目的について説明し、課題とそれに対するアプローチを述べる。まず自然言語処理で用いられる電子化辞書や従来の伝統的な類義語辞書などの適用分野やその辞書構造について概観する。その上で本研究の取り組む課題として、コーパスの解析に基づく語の相互関係の発見を取り上げる。そして、最後に本論文の全体構成について述べる。

第２章では、本研究に関連する研究として、従来の類義語集合や階層関係の抽出手法の概観を行い、本研究の位置付けを明確にする。

第３章では、同じ観点で類義関係にある語をグルーピングする方法について議論する。

ここでは、コーパス中に含まれる複合名詞から修飾関係を抽出し、それらによる語と修飾関係を頂点と辺とするグラフ構造中から極大完全2部グラフ部分を類義語集合として探索し抽出する手法を提案した。評価実験では1ヶ月分の新聞記事コーパスから得られる修飾関係をグラフ化し、その中に含まれる極大完全2部グラフ（約4900個）を抽出した。得られた類義語集合の正解判定は人手で行い、その結果、２頂点同士からなる2部グラフにおける正解率は約30%であった。また観点の違いによる類義語集合の獲得に関しては、同じ語に対して違う観点で複数の類義語集合が得られることを確認した。これにより本研究の第一の課題に対し提案手法が有効であることを確認した。

第４章では、多義性を持つ語の発見方法について議論する。語の相互関係を考える場合、同じ語でもその語義ごとに分けて扱わなければならない。そこで本研究ではこの課題を解決するために、多義性を持つ語の発見方法を提案する。ここでは、多義語となる語はその語義によって異なる類義語集合に含まれることに注目する。そこで、本手法ではまず３章の手法で類義語集合を抽出し、次に多義性を調べたい語について、その語を含む類義語集合の数を調べて複数の類義語集合に属するなら多義性を持つと判断する。ところがコーパス中の修飾関係が少ない場合、同じ観点による類義語集合が複数得られるという問題がある。そこで、本手法は構成要素が類似した類義語集合は併合する手順を加えてこれに対処している。評価実験では1年分の新聞記事コーパスから得られる修飾関係をグラフ化し、極大完全2部グラフ（約1,300,000個）を抽出した。この実験では抽出グラフ数が多いため、高速なグラフ探索列挙アルゴリズムとして逆探索法を用いた。次に得られたグラフに併合操作を行い、形態素解析誤りなどを含むグラフを除去した後、複数の類義語集合に含まれる語を集計した（182語）。そしてこれらの語の多義性を人手で判定した。その結果、そのうち31語については多義語であることが確認された。これにより、多義語の発見手法の有効性を確認した。

第５章では、IS-A関係の発見方法について議論する。ここでは動詞と名詞の依存関係を用いた発見方法を提案する。まずIS-A関係の定義を明確にし、これよりIS-A関係の発見手法を導いた。具体的には動詞と名詞の依存関係を解析し、名詞ごとの動詞集合を作成し、それらの包含を比較して、与えられた名詞の下位語候補となる順序リストを作成する手法を提案した。評価実験ではまず11年分の新聞記事コーパスを解析し、名詞ごとの動詞集合を作成した。次に評価に用いる上位語のサンプルリストを分類語彙表から作成した。そしてこのリストに含まれる名詞に対して提案手法を用いて下位語候補を作成した。この候補語リストの正解判定は人手で行なった。その結果、平均正解率は約34%であった。また従来手法とのカバレッジの違いを確かめるために、同じコーパスから定型表現を用いてIS-A関係を抽出し、本手法によって得られたIS-A関係と比較した。その結果、本手法で得られるIS-A関係の約6%が従来手法によって獲得されることを確認した。これにより、本手法の有効性が示された。

第６章では、本研究の結論と将来の展望について述べる。