審 査 の 結 果 の 要 旨

氏名 ステネートルプ ポントス　ラース　エリック

自然言語処理において，未知語は常に問題となる．近年急速に研究の進んだ機械学習に基づく自然言語処理において，いかに大規模な辞書やコーパスを学習に用いても，学習時に含まれない語彙が運用時に観測されることは避けられない．この問題の解決への糸口にはいくつかの可能性があるが，そのひとつとして挙げられるのが，字句表現（綴り）の特徴である．たとえば「人名らしい特徴を持つ綴り」や「地名らしい特徴を持つ綴り」という類の特徴である．そのような特徴は，それだけで人名か地名かを判断できる決定的な情報にはなりえないが，機械学習に基づく自然言語処理では貴重な情報となりうることが期待される．

本論文で論ずるのは，辞書や大規模コーパスを入力として機械学習を適用することにより得られる，字句表現の特徴と対象物の特徴との相関である．これを本論文では「語彙的な事前知識」と呼んでいる．この事前知識を未知語等に適用することにより，言語的な不確定性を減少させ，自然言語処理に活用することを目指す．

本論文は以下のような7章からなっている．

第1章では，自然言語処理における未知語の問題を提起している．この問題は生物医学文献の領域ではとりわけ顕著で，同じサイズのコーパスに含まれる語彙の数が，新聞記事の場合のおよそ2倍に達することが指摘されている．

第2章では，まず最初に，未知語の大きな部分を占める固有表現を文中から取り出す「固有表現抽出タスク」における辞書の有用性が，既存知識として導入されている．しかしながら既存手法の多くは辞書語との完全一致に基づいており，表記の揺れや字句表現の特徴を考慮するものではない．続いて，SimSemと呼ばれる手法が提案される．これは近年提案された近似文字列照合アルゴリズムを用いて新たな素性集合を生成する．この手法が狙っているのは，同じ対象を表す単語にも複数の表現上の差異が認められる点に着目し，そのような表現上の多様性を考慮することのできる素性集合である．提案手法は，生物医学文献における意味的カテゴリの曖昧性解消タスクを用いて評価されている．このタスクは，文章中の指定した固有表現の意味カテゴリを判定するものである．6種類の生物医学文献コーパスを用いた実験のうち，CALBC CIIにおいて，提案手法は既存手法を明らかに上回る性能を達成し，その他のコーパスでも既存手法に劣らないことが示された．

第3章では，第2章の手法のチューニングが論じられている．まず，近似文字列照合に対して EDIT とそれを文字列長で正規化した NEDIT の2種類を考察している．また文字列の先頭と末尾を示す特殊記号の利用が提案される．また，言語リソースの選択的使用，閾値のチューニングが論じられている．第2章と同じ評価を行い，性能の向上を確認している．

第4章では，提案手法を生物医学から一般分野のテキストに適用し，9 つの異なるコーパスで評価している．ここでは候補カテゴリを絞り込む曖昧性解消タスクを想定し，曖昧性の低下と再現率の維持をバランスさせる指標を提案し，評価に用いている．その結果，提案手法は 98% から 99% というほぼ完全な再現率を維持したまま，曖昧性を劇的に削減することができることを示した．

第5章では，提案手法の自然言語テキストに対するアノテーション作業支援ツールへの応用が述べられている．自然言語処理の研究のためには人間が定める「真値」が必要であり，そのために多くの人的リソースが割かれている．本章では生物医学文献へのアノテーション作業ツールに対して，提案手法によりアノテーション候補を絞り込むことで，アノテーション作業者の作業効率の向上が達成されることを実証した．

第6章では，語彙的な事前知識として語義の低次元空間上へのマッピングを用いる手法に注目し，語義の獲得に用いる外部リソースの分野やサイズの影響について調査している．生物医学文献における固有表現抽出タスクおよび曖昧性解消タスクを評価基準として，新聞記事および生物医学文献に基づくブラウンクラスタ，Google N-gram クラスタ等の手法を取り上げ，性能を評価した．その結果，固有表現抽出タスクでは生物医学文献による学習結果が一般文献による学習結果よりもよい成績を挙げる傾向が見られたが，曖昧性解消タスクでは逆に Google N-gram クラスタが高い性能を発揮した．

第7章は本論文の知見を総括している．

上記のように，本研究は自然言語処理における語彙的な事前知識の有効性を様々な視点から明らかにするものとなっている．語彙的な事前知識の活用はこれからさらに広がってゆくものと期待されるが，本論文の成果はその先端を開拓する優れた研究成果であり，今後の自然言語処理技術の革新的進歩，すなわち情報技術による人類の知識の集約と活用とに大きく貢献することが期待される．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。