平成19年度 情報工学コース卒業研究報告要旨

石川 研究室	氏　　名	酒 井　佑 太
卒業研究題目	大規模テキストからの 知識獲得のための文分割に関する研究	

インターネットの普及により世の中に大量の情報があふれるようになった．しかし，それらの情報は雑多に存在しているため，有効に活用する事が困難な状況である．情報を有用に扱うためには，情報に内在する知識が適切に構造化されている必要がある．世の中の知識の多くはテキストで記されており，大量のテキストから知識を抽出することが，有用であると考えられる．

本論文では，テキストから知識データベースを生成することを目的に，文から知識を抽出する手法を提案する．本手法では，知識の単位として単文を想定している．単文とは述語を一つ含む文である．言語処理技術の進歩により，単文であれば様々な処理が高精度に行える．知識の単位に単文を用いることにより汎用的な知識データベースの構築が期待できる．本手法は，重文や複文を複数の単文に分割することにより知識を獲得する．本手法における知識抽出の例を図 1 に示す．

文分割には節境界情報を利用した．節とは一つの述語を中心とするまとまりである．毎日新聞コーパス中の 13 記事 102 文に対して，節境界での分割を行った．節境界の検出には節境界検出プログラム CBAP を用いた．分割結果の分析を行った．分析には京都テキストコーパスの構文情報を用いた．京都テキストコーパスは毎日新聞コーパスに構文情報を人手で付与したものである．分析の結果，以下の問題が確認された．

1. CBAP は，厳密には節境界ではない統語的な切れ目に対しても境界を検出するため，述語を含まない非文が抽出される．
2. 他の節との意味的な結びつきが強く，本来の意味が失われる節が存在する．
3. 文節間の係り受け関係が節境界をまたいでいる場合，係り受け関係が切断され，本来の意味が失われる．

これらの問題を解決するために，それぞれ以下の方法を導入した．

1. 抽出された非文と，意味的に結びつく節とを結合する．
2. 分割された単文に文番号を付与し，意味的結びつきが強い節とその文番号とを結合する．
3. 係り受け関係が節境界をまたぐ文節の直後で分割を行い，意味的に結びつく節と結合する．

意味的に結びつく節の検出には構文情報を利用した．図 2 に本手法による文分割の例を示す．

京都テキストコーパス中の 33 記事 465 文に対して，評価実験を行った結果，83%という高い正解率を達成し，本手法の利用可能性を確認した．

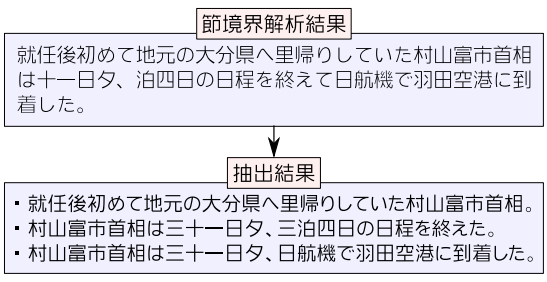

図1 テキストからの知識抽出の例

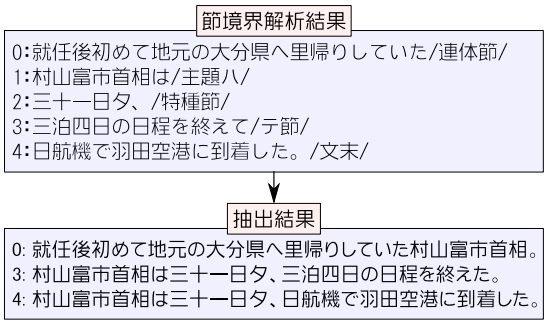

図2 本手法による文分割の例