インドネシア語の形態素解析に関する研究

Muizzul Hidayat

北陸先端科学技術大学院大学 情報科学研究科

1997 年 5 月 15 日

キーワード： インドネシア語, 形態素解析, 言語知識の獲得.

形態素解析は自然言語処理において重要な処理の一つである．形態素解析の主な目的は与えられた文を形態素・語の並びに分割し，それぞれの形態素・語の品詞などを決定することである．また，自然言語処理におけて形態素解析は第一段階の処理であり，他の処理（統語解析，意味解析，談話解析等）に有用な情報を与える役割を担っている．

これまで形態素解析に関する研究は様々な言語に対して盛んに行なわれてきた．しかし，英語や日本語の形態素解析の研究と比較して，現在インドネシア語形態素解析はまだ十分に研究されていない．その一つの原因はインドネシア辞書の電子化はかなり遅れており，既存の辞書ではまだ十分にインドネシア単語をカバーしていないからである．また，インドネシア語は他の言語にあまり見かけない特徴を持っており，これらの特徴を把握しそれにあった処理体系を築く必要がある．

そこで，本研究ではインドネシア語の特徴を研究調査し，既存の辞書からより適応範囲の広い辞書を構築し，その有用性を実験的に確認することを目的とする．インドネシア語の単語は大きく分けて基語と構成語がある．ここで，基語とはこれ以上分割できない単語のことであり，構成語は基語の反復語（重複語）や基語（または重複語，構成語）に様々な接頭辞や接尾辞が付加されて作られる単語等のことである．このようにインドネシア語の語構成に関する規則は，かなり複雑である．また，接頭辞や接尾辞はほとんどの基語や構成語に付加でき，その組合せが膨大なインドネシア語の単語を生成する．

一方，インドネシア語の形態素解析の実現においては，この構成語をどう取り扱うかが問題になってくる．インドネシア語の形態素解析を実現する方法は，大きく２つの方法に大別できる．第一の方法は，全ての単語を辞書に登録しておく方法である．この方法をとる場合，形態素解析のアルゴリズムは単なる辞書引きでよく，非常に簡単になる．しかし上述のようにインドネシア語では，接頭辞，接尾辞および基語の組合せが膨大の数にある．この膨大な数の単語をどうやって網羅的に収集するかということが問題になってくる．

第二の方法は，基本となる単語（基語）だけを辞書に登録しておき，語構成に関する規則（語構成規則）を併用する方法である．この方法では，形態素解析のアルゴリズムは多少複雑になるが，辞書のサイズは小さく押えることができる点で有利である．しかしこの方法では，語構成規則が必要である．これらの規則を人手で作成することは可能であるが，客観性が欠けているという問題がある．

そこで，本研究は第二の方法の問題を解決するために，既存の辞書から語構成規則の自動獲得を検討する．

ここで用いる辞書は，CICC プロジェクトで作成された IMD（Indonesian Master Dictionary）である．この辞書はすべての単語を登録するという立場に立つため，基語の他に構成語も含んでいる．まず，辞書中に含まれる構成語を分解し，接辞のリストと語構成規則を獲得する．次に，こうして得られた知識を形態素解析時に利用し，ＩＭＤに含まれていない単語も解析できる，より広いカバレジを持った形態素解析システムを実現する．

本論文の構成は，以下の通りである．まず，第２章で，インドネシア語の形態素の特徴と，インドネシア語の形態素解析の現状について述べる．第３章では，ＩＭＤから接辞リストや語構成規則をいかにして獲得するかについて述べる．第４章では，本研究で作成した形態素解析システムについて述べ，第５章では，実験と評価について述べる．最後に第６章で，結論を述べる．