武　田　浩　一（黒橋教授）
「Building Natural Language Processing Applications Using Descriptive Models」
（記述的モデルを利用した自然言語処理アプリケーション構築）
平成 22 年 3 月 23 日授与

デジタル化されたテキスト情報の氾濫により、自然言語処理技術に基づいたアプリケーションへの要求が急速に増大している。特にインターネット上の Web ページやブログなどに含まれるテキスト情報の検索や翻訳をはじめとしたアプリケーションは、インターネット上の豊富な情報やサービスを統合し、多くの利用者にとって欠かせない社会的インフラを構成するようになった。さらにアプリケーションの多様性は、形態素解析、構文解析、機械翻訳のための構造変換、情報抽出といった要素技術のコンポーネント化を促進した。このため各コンポーネントがルールベースあるいは機械学習的手法といった実装の違いにかかわらずコンポーネント間のインタフェースを容易にするためには、各コンポーネントの入出力を明確にする自然言語処理の記述的モデルを積極的に利用することが極めて重要になった。

本論文は、このような記述的モデルに基づいて形態素解析の 1 つの主要機能である漢字複合語分割、機械翻訳（概念表現を利用した中間言語方式と同期文法を利用したパターン翻訳方式）、中間言語表現の言換え、およびテキストマイニングの一手法である集合知の意見集約、という 4 種類のコンポーネントを実現した研究をまとめたものであり、以下の主要な成果を得ている。

1. 漢字複合語の造語モデルをマルコフモデルによって表現し、大量のコーパスから各漢字複合語が生起する確率を学習するとともに、Viterbi 法によって与えられた漢字複合語の最も尤度の高い分割を推定するアルゴリズムを提案した。また同アルゴリズムにより、電気工学分野の技術論文からランダムに抽出された長さ 3 文字以上の漢字複合語に対して 95.0% の平均分割精度を達成することができた。図 1 に長さ 9 文字の漢字複合語の最尤分割例を示す。各行には与えられた漢字複合語、分割パターン（P は接頭辞あるいは漢字複合語に前接する 1 文字漢字、1 および 2 はそれぞれ 2 文字漢字複合語の 1 文字目および 2 文字目、S は接尾辞または漢字複合語に後接する 1 文字漢字を示す）、およびその生起確率が対応している。

一様振幅周波数特性	121212S12	-1.27093E+01
一端子周波数依存性	12S12S12S	-1.43907E+01
一般化方程式誤差法	12S12S12S	-1.24734E+01
一般化相互相関器法	12S1212SS	-1.22991E+01
一般化可到連性空間	12SP12S12	-1.45771E+01
一般化優度比統計量	12S12S12S	-1.64295E+01
一般化反作用原理及	12S12S12S	-1.53643E+01
一般化逆固有値問題	12SS12S12	-1.43205E+01
一般産業用電気機器	1212S12SS	-1.31461E+01
一般産業用電気設備	1212S1212	-1.38083E+01
一般的多領域信頼性	12SP1212S	-1.28993E+01
一般的感度解析問題	12S121212	-1.09410E+01
一般料金改正用料金	121212S12	-1.70158E+01
一般形伝達関数実現	12S121212	-1.27132E+01
一階連立微分方程式	12121212S	-1.83198E+01

図 1：長さ 9 文字の漢字複合語分割例

2. 中間言語方式の機械翻訳において概念トランスファーという手法により原言語および目的言語における概念表現の差異を吸収するとともに、集合や変数に相当する概念を導入し、並列句や代名詞などの広範な自然言語表現を実用的な精度で翻訳できるシステムを試作した。また「色」属性をもつ有向グラフに基づく概念表現の言換え手法による、目的言語の制約のもとで文生成が可能となる概念表現が計算できることを示した。
3. 同期文法によるパターン翻訳方式を提案し、インターネットのオープンドメインな Web ページの翻訳に適した英日翻訳システムを実装した。同期文法中の非終端記号に語幹や対応関係の記述を許す拡張を行うとともに、その数学的性質を明らかにした。また、このシステムは製品化され外部からも高い評価を得た。
4. インターネットにおける掲示板のように意見対立が存在する文書集合に対し、意見集約のための計算手法を提案した。論理的な矛盾に基づく意見のグループ化の計算的複雑さ（NP 完全）を示すとともに、3 分木を利用した対立意見の集約手法を示した。

実用的な自然言語処理アプリケーションを構築するうえで、既存の知識や資源を再利用し、漸進的に機能を拡張し、多様性に対応することは産業的な要請である。このために、できる限り汎用的なコンポーネントを設計し、記述的モデルを通して理論的な諸性質を明らかにするということは大きな目標であった。

今後は人間の感情表現の扱いや質問応答といった自然言語処理の関連分野も考慮し、記述的モデルを適用して意味的により豊富なテキスト情報を処理できるような研究に取り組みたい。