福本　文代
山梨大学大学院医学工学総合研究部　助教授

大規模コンテンツからの知識獲得と情報検索への適用

近年、世界規模の情報ネットワーク構築が進むにつれ利用可能な電子化文書の多言語化が進んでいる。 機械翻訳システムは自然言語処理の成果の一つであり、多言語の問題に対処するために、研究・開発・商品化された。しかし現在の翻訳技術は、残念ながら、言語表現の多様性に十分対処可能な枠組みを提供しているとは言えず、インターネットを介してユーザが望む情報を正確に検索・提示するためのコア技術として利用するには不十分であると言わざるを得ない。現在の翻訳技術を多言語情報検索に利用するためには大量の言語知識が必要であり、検索対象となる大規模データから構文や意味といった翻訳知識を高精度で抽出する技術が必要不可欠である。80 年代までの知識獲得研究は特定の現象を深く解析するか、大規模な文法や辞書を手作業で作成するかいずれかの手法を採用する場合が多かった。しかしいずれの手法も様々な領域の現実的な言語現象に対処するには困難が伴う。90 年代になり、電子化されたデータから統計や機械学習を用いて言語知識を自動的に獲得する研究が盛んに行われるようになった。例えば形態素解析のためのオートマトンモデルの学習、文法規則の学習、また最近では高精度でパラメータが学習できる機械学習も提案され、1 万規模のデータに対してその有効性が示されている。しかし現実の多様な言語現象に耐え得るだけの知識を獲得するためには、より大規模なデータに対しても精度・計算量の点で処理可能な枠組みを構築することが必要である。本研究では、大規模 Web コンテンツから、多言語を対象とした情報検索に必要となる対訳知識を自動的に抽出する手法を確立することを目的とする。

研究成果の発表

Using Category Hierarchies for Correcting Category Errors in Multi-labeled Data
Proc. of Human Language Technologies as a Challenge for Computer Science and Linguistics　P211-215, 2005

Generating Category Hierarchy for Classifying Large Corpora
IEICE TRANS. P1543-1554, Vol.E89-D, No.4, 2006

分野の階層構造を利用したコーパスの誤り修正と文書分類への適用

電子情報通信学会論文誌，P552-566，Vol. J89-D No.3，2006