審 査 の 結 果 の 要 旨

氏　　名　　ホリエ　アンドレ　ケンジ

本論文は、「Interlingual Semantic Analysis of Text: Alternative methods to full corpus annotation (テキストの中間言語方式意味的解析：全コーパスアノテーションの代替法)」と題し、英文で５章により構成されている。

近年、自動翻訳には、大きく２つの方式が研究されている。一つは統計的翻訳方式で、翻訳対集合を大量に集めることで元言語から対象言語に直接翻訳する方法であり、もう一つは中間言語方式で、言語に依存しない形式的な中間言語 (interlingua) を定義し、元言語をそれに翻訳し、中間言語から対象言語に翻訳する方法である。

一般に、中間言語方式は、汎用の中間言語を定義することが難しく、中間言語への翻訳も難しい。このため、近年は統計的翻訳が主流となっているが、統計的翻訳は文の対応のみを利用する以上、解析は浅く文の理解は行わず、その意味で処理結果には限界があると言わざるを得ない。これに対して、文の意味を深く解析することで、より正確な翻訳結果を得る事ができる可能性がある。そこで本論文では、中間言語を汎用に考えるのではなく、すべての言語に普遍に見られる言語のある側面として、態や時制等に限って中間的な記述の枠組みを用意し、これを用いて文を解析することを提案している。解析にあたっては、教師有り学習を用いているため、解析のための学習データを手動で作成するコストをいかに抑えるかが問題となり、このコストを抑えるような中間的な記述方式を提案している点に本論文の特徴がある。

第１章は「Introduction」であり、関連分野の現状と課題をまとめ、本論文の目的を述べている。

第２章は「Contextual Semantic Relations」と題し、文脈に基づく単語や文部分や意味的関係を述べている。コーパスのアノテーションに基づいて処理を行う際の問題は、アノテーションのクラスの統計的な偏りである。クラスの中には、少数回しか現れないものがあり、これを原因として機械学習における性能の限界が生じる。ここでは、この問題に対処するために、ブートストラップ法を用いて集合の拡張を行う方法を提案し、特に、特徴量の距離に基づく弱い学習器を構築することで、ブートストラップ処理が初期的には観測不十分であったとしても、webから類似の事例候補を探し、雑音を除きながら、もとの少数の事例と同じクラスかどうかをアノテーション作業者に確認してもらうことで、より大きな学習データを構築することで、最終的には高い性能を得る方法を提案している。またこの方法を用いることで、単純に特徴量ベクトルを用いる時よりも、

より高い適合率を実現できることが実験的に示されている。特に頻度の低いクラスに関しては、汎用のアノテーションを行う場合よりも、アノテーションのコストを大幅に下げルことができることを示している。

第３章は「Modality」と題し、態に関して述べている。本論文では、文の態と態表現の相関が高いことから、文の態のアノテーションがあれば、態表現を得ることができると仮定して、態表現をアノテーションしなくとも、異なる言語の間で対応する文の集合が得られるならば、ある言語で態がわかっている時にはおのずと対応する文の態もわかることになり、言語に依存性が高い態表現のアノテーションを行う代わりに、文の態のアノテーションだけに依拠して、態表現を自動で得て文の態の判定に利用する方法を提案している。この目的のために、本論文では、文の態がわかっている時に、態表現を得る解析方式を提案し、実験的にその有効性を示している。しかも、アノテーションに関しては、通常は文の中を人間が解析して、態を決める表現にアノテーションをしなければならなかったのが、本論文では文単位で態のアノテーションを行うだけでよい上、文の態が一端わかると、それと対応する別の言語の文の態も判定できるという利点がある。

第４章は「Tense」と題し、時制について述べている。時制に関するまとまった研究は、言語学上も自然言語処理の上でも少ない中で、本論文では、言語学において提案されているReichenbachの枠組みを元に研究を行っている。この枠組みは言語汎用の考え方で時制を記述するものであり、本論文では、これをもとに文の時制のアノテーションの新しい方法を提案している。この方法を用いると、文の時制に関する部分のアノテーションさえあれば、そこから文全体の時制を推論して得ることができる。また、文の部分に付与するアノテーションは時間という一軸に沿った論理的で自然なものとなっており、作業者の直感に沿うものとなっている。

本論文では、アノテーションが付けられた文集合がある時、そこから文の時制を推定するアルゴリズムを併せて提案している。論理的でわかりやすいルール群を定義し、これを元に既存の教師有り構文解析手法を応用することで実現している。実際に文の時制の推論を行い、その有効性を実証している。

第５章は「Discussions and Conclusion」であり、本論文の成果をまとめるとともに、関連技術の将来の方向性を論じている。

本論文ではすべての言語に共通する言語的な側面の観点から、文脈の観点からの文部分の意味的関係とともに態や時制を捉え、それを丁寧に自動解析しようとすることで、現状の統計的翻訳あるいは広く自然言語処理では扱い切れていない部分に光を当てている。そこで問題となることは、教師有り学習として解析を実現する上でのアノテーションのコストである。本論文では、アノテーションの記述体系自体を工夫することで大幅にコストを削減できることを示すとともに、時制や態といった自然言語処理においてはそれほど扱われてこなかった言語的な側面を丁寧に解析し、実際にどの程度の改良ができるのかを示している。すなわち、本論文の成果は、自然言語処理に貢献するだけで

なく、理論言語学、さらには応用言語学やその工学的応用において、関連分野の発展に貢献するとともに、情報理工学における創造的実践の観点からの価値が認められる。

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。