氏名	菊井玄一郎（きくいげんいちろう）
学位(専攻分野)	博士（情報学）
学位記番号	論情博第76号
学位授与の日付	平成19年3月23日
学位授与の要件	学位規則第4条第2項該当
学位論文題目	Corpus-based Ambiguity Resolution for Cross-language Information Processing （クロスランゲージ情報処理のためのコーパスに基づく曖昧性解消）
論文調査委員	（主査）教授 河原達也　教授 奥乃博　教授 黒橋禎夫

論文内容の要旨

本論文は，自動翻訳や言語横断検索などのクロスランゲージ情報処理を対象として，コーパス（テキストや音声のデータ）から言語的知識を取り出すことにより，これらの処理過程で出現する曖昧性を解消する方法に関する研究をまとめたものである。コーパスを用いた曖昧性解消においては，対象タスクの言語表現を収集し，必要に応じて対訳を付与するなどコーパス構築のコストが問題となる。本論文では，自動翻訳における訳語の曖昧性，およびインターネットテキストや音声をクロスランゲージ処理するためのテキストに変換する際に生じる曖昧性を取り上げ，コストの小さいコーパスを用いて曖昧性を解消する方法を提案し，有効性を示している。

第１章では，本研究の位置づけを述べるとともに，本論文で取り上げる主要な問題とアプローチを述べている。

第２章では，言語横断検索や自動翻訳で必要不可欠な内容語（キーワード）リストの翻訳において，翻訳先言語の単言語コーパスのみを用いて，訳語の曖昧性を解消する方法を提案している。提案手法では，リスト中の各単語に対する訳語候補の中で，最も類似した文脈に出現しうる組み合わせを選択する。単語間の文脈の類似性は，共起行列を次元圧縮したベクトル空間におけるコサイン距離で評価する。評価実験の結果，既存の手法に比べて最大で13.8%の翻訳精度の向上により，提案手法の有効性を示している。

第３章では，同一分野ではあるが対訳関係にない２言語コーパス（対照コーパス）を用いて，前章と同様の訳語の曖昧性を解消する手法を提案している。まず前処理として，翻訳元コーパスにおいて翻訳対象単語の前後に出現する単語列（文脈）を収集してクラスタリングし，２章の方法を用いて各クラスタに対して１つの訳語を対応付ける。翻訳時には内容語リストを文脈とみなし，各単語に対してこの文脈に最も意味的に近いクラスタを選び，対応する訳語を出力する。実験の結果，前章の方法に比べて4.3%の翻訳精度の向上により，２言語のコーパスを使うことの効果を確認している。

第４章では，インターネット上のテキスト符号列に対して，符号系および言語を識別する手法を提案している。本手法では，適用可能な全ての符号系を用いてテキストに復号した後，得られた各テキストの言語的尤度を各言語の統計モデルによって評価し，最も尤度の高い符号系および言語を選択する。実験の結果，符号系の識別精度99.7%（識別対象は7bit符号系３個と8bit符号系５個の計８符号系），言語識別精度95.2%を達成した。

第５章では，音声翻訳を対象として，音声認識で生じる曖昧性に対処する手法を提案している。本手法では，音声認識結果の上位Ｎ個の候補全てに翻訳処理を適用し，音声認識で得られる音響的・言語的スコアと翻訳で得られるスコア群（訳語の対応度や翻訳結果の言語尤度等）を対数線形モデルで統合することにより，最適な結果を選択する。ここで，各スコアの重みは，対話音声を書き起こして正解訳を付与したコーパスにより最適化する。提案手法により，認識候補を１つのみ用いる通常の方法に比べて，音声認識精度および翻訳精度の両方で改善を実現している。

第６章では，特に整備の難しい学習コーパスとして，５章で用いた音声翻訳用コーパスの構築法について検討している。ここでは，会話調の表現を作文することにより作成したコーパスと単言語の対話コーパスを混合することにより，音声翻訳システムを介した模擬対話が近似できることを定量的に明らかにし，これに基づくコーパス構築方法について論じている。

第７章では，提案手法の効果や今後の課題について論じながら，結論を述べている。

論 文 審 査 の 結 果 の 要 旨

本論文は，自動翻訳や言語横断検索などのクロスランゲージ（多言語）情報処理において生じる，訳語選択の曖昧性と入力系に起因する曖昧性を解消する方法に関する研究をまとめたものであり，得られた主な成果は次の通りである。

１．単語の多義性に起因する訳語選択の曖昧性を解消する２つの方法を考案した。両手法とも，単語の文脈をモデル化することにより，文脈に最も沿った訳語を選択するものであるが，翻訳先言語コーパスにおける共起関係に基づく部分空間上で類似度を定義する手法と，翻訳元言語コーパスで文脈のクラスタリングを行い，訳語を対応づける手法である。いずれも作成コストの高い対訳コーパスを必要としないもので，適用可能性が大きいと考えられる。

２．言語や符号系が未知のテキスト（符号列）に対して，これらを自動識別する方法を考案した。提案手法では，すべての符号系を用いて復号を試み，統計的言語モデルによる尤度によって判別を行うもので，コーパスがあれば，規則を全く記述する必要はなく，しかも99.7%の符号系識別精度を実現している。

３．音声認識システムが出力する誤りを含む単語列を対象として，自動翻訳を改善する方法を考案した。提案手法は，音声認識結果の複数候補に対して，それぞれ翻訳を試み，翻訳の尤度も考慮して最適な候補を選択するもので，音声認識精度と翻訳精度の両方で改善を実現している。また，自動音声認識と機械翻訳を密結合した先駆的な研究と位置づけられる。

以上のように本論文は，自然言語処理において重要なテーマの１つである曖昧性解消の問題に関して新たな知見を与えるとともに，クロスランゲージ処理において有用な方法論を提示するもので，学術上・実際上寄与するところが少なくない。よって，本論文は博士（情報学）の学位論文として価値あるものと認める。

また，平成19年１月15日実施した論文内容とそれに関連した試問の結果合格と認めた。