成果報告書

Ⅰ．研究概要

氏名	蔡宜静（サイ ギセイ）
所属	台湾康寧大学応用外国語学科　准教授
招聘回（招聘期間）	1回（2013年4月1日～2013年9月30日）
招聘研究テーマ	「英日中韓言語を対象とした言語依存しないテキストデータ分析に関する研究」
研究目的	現在までの研究成果では、文字*N*-gram法という技術を駆使した言語固有の文法知識を一切利用しない言語体系から独立した同一の方式を提案し、英語、日本語、中国語の電子的な新聞記事を分類させるという実験を行った。今回は、日本語と中国語の言語系統にともに近い韓国語を導入し、この操作によって、複数以上の言語に跨るテキスト自動分類の方法に寄与したい。

研究概要：

今回の研究対象は、英語の「ロイター通信」(計12902件)、日本語の「毎日新聞」(計10500件)、中国語の「人民日報」(計9000件)、韓国語の「ハンギョレ」(計7500件）という四種の電子新聞記事である。本提案手法の数理モデルを用い、四種の異なる言語体系のテキストにおいて、それぞれの分類精度の数値がどのように変化するかを検証し、その意味付けを行った。従来の研究手法と違い、本提案の手法は、最初の段階から文書に形態素解析をかけるプロセスを採用しないため、言語依存しないという点に第一の特徴がある。テキスト自動分類には、①学習段階と②検証段階という二つのステップがあります。学習段階からカテゴリー付きのトレーニング用の文章を使い、効果的なタームを抽出し、コンピューターに学習させながら、タームデータセットに保存する。この時点から、蓄積手法を導入する。そして、検証段階において、カテゴリーを伏せたテスト用の新しい文書に同じく情報処理過程を行い、カテゴリーの自動分類と文書分類の結果を検証する。この二つのプロセスを繰り返し、各言語のテキストの分類精度を検証する。

本研究の第二の特徴として、二段階における各カテゴリーの条件付き確率の計算は、乗算ではなく、加算を用いるため、大幅に計算量を軽減できる。そして、文字*N*-gramのパラメータ値を順番に変化させることにより、各言語に最適の*N*値を自動的に見つけ出すことが可能である。

各言語の最良精度の実験結果は、英語の最良精度が文字6-gramの94.5％、日本語の最良精度が文字5-gramの88.9％、中国語の最良精度が文字4-gramの92.6%、韓国語の最良精度が文字4-gramの90.6%であった。各言語によって最適なN値が異なる。これは、各言語の一文字における情報量の多少との関連があるかどうかについては、今後の検討課題である。ともかく、この提案手法は、文字*N*-gramの*N*値を変化させることにより、対象文書の分類難易度を自動的に識別することが可能である。そのため、提案手法の発想は単純なものであるが、テキスト自動分類の際、有効なだけではなく、非常に高い分類性能が得られた。

展望：

本研究成果の応用可能性への展望は、現時点において以下の四つの方向から着目している。

1.　　他言語への適用： 異なる言語系統や適用言語の拡張(台湾諸語、古典語など)。

今回用いた英語、日本語、中国語、韓国語の四種の言語テキストは、それぞれインド・ヨーロッパ語族、ウラル・アルタイ語族、シナ・チベット語族、ウラル・アルタイ語族に属しているので、異なる言語体系の比較だといえる。今後はほかの言語系統のテキストを使ってもいいし、または同一祖語から生じた同系統の諸言語だけのテキストを見比べてもいいだろう。あるいは、中古日本語と漢文文語文のテキスト分類を検証したいときも、本研究の手法が利用できる。

2.　　応用事例の拡張：

経営情報サービスの場合は、たとえば、消費者アンケートの分類に使うことが可能である。教育分野の場合は、たとえば、文系と理系のレポートの自動分類によるクラス編成へのフィードバックなどに生かすことも可能である。

3.　　基礎研究：

単純な状態遷移モデルにもとづき、形態素解説から言語独立の単語を切り出す方式を、今後の方向として検討していく予定である。 私どもの提案した文字 *N*-gram は、言語固有の文法知識は一切利用しないが、単語 *N*-gram の手法を使う場合は、やはり初期段階において、形態素解析のプロセスを通過しないと実験できないという側面もある。そこで、「状態遷移モデル」をこれから開発しなければならない。現段階は、形態素等とは一対一に対応しないような手法に関する提案を討論しているところである。この部分の基礎研究ができると、単語 *N*-gram の実験の改良が可能になる。

4.　　応用研究：

本提案手法と言語学のコーパス研究の成果との関連について考え、また、その差異を比べて、言語教育に提携できるような共同研究を今後の検討課題としている。そもそも言語研究のためのコーパスを利用する際、研究者自身も対象言語に関する記述文法、言語理論などに対する知識が必要である。本研究は、言語依存していないので、言語研究のためのコーパスの研究成果との関連、または、差異を明らかにすることができれば、新たな応用研究の可能性が開けるのではないかと考えられる。その橋渡しの方法について、今後の研究課題として考えていきたいと思う。さらに、自動分類の手法に、テキストマイニングの自動要約の手法と併用し、講義要約文と講義理解などの研究テーマの開発も、今後の新たな研究課題として考えていきたいと思う。