統計的機械学習法による多言語間テキスト要約

Minh Le Nguyen
情報科学研究科
北陸先端科学技術大学院大学
2004年7月

テキスト自動要約は，与えられたテキスト文書から最も重要な文脈を生成するタスクであり，ユーザやアプリケーションの要求を反映させながら元の文書を短縮するタスクである．テキスト自動要約翻訳（cross-language text summa-rization: CLTS）は，与えられた文書を元の言語とは異なる言語で要約するタスクであり，単一言語のテキスト自動要約とは異なるものである．近年のテキスト情報量の急速な増加や，特定言語に不慣れなユーザの増加により，CLTS タスクは多くの注目を浴びている．CLTS システムは，本質的には，テキスト要約 と機械翻訳エンジンの組み合わせである．もし機械翻訳システムとテキスト要約システムが十分に高性能ならば，CLTS システムはそれらの組み合わせにより 簡単に実現可能である．しかしながら，そららのシステムを得ることは切望さ れているにも関わらず困難な作業であり，特に，ベトナム語などのように英語 や日本語以外の言語に対してはさらに困難な作業になる．それゆえ，CLTS に関する研究は非常に難しいとされている．本論文の目的は，精度が高く，計算コストの低い CLTS システムを実現することである．そのために，テキスト要約の専門家の行動に基いた統計的機械学習法 に注目し，これを CLTS に適用する．学習法を適用することの利点は，言語知識 を構築する際に，最小限の作業で高い精度を実現可能なことである．提案するシステムは，重要文抽出，文圧縮，翻訳の 3 つの主要タスクからなる．最初に，文書の要旨に最も関連している重要文の集合を文書全体から抽出し，次に，抽出した重要文をそれらの意味を保ったままより短い文へと圧縮する．最後に，それらの文を翻訳し，異なる言語の要約を得る．重要文抽出に関しては，コーパスに基づく抽出法を調査する．さらに，最大エントロピーモデルに 基づいた共学習法を研究し，ラベル付けされていないデータを利用することで 文抽出の性能を向上させる．評価実験により，機械学習を使用する重要文抽出 にはラベル付けされていないデータが非常に有用であり，共学習法が適することを示す．

長文を短文に圧縮する文圧縮に関しては，長文の構文木を小さな構文木に変形 する処理として公式化する．この処理は木を小型化するための一連のアクショ ンからなり，小型化された構文木から簡単に圧縮文を得ることが可能である．ここで重要なことは，構文木に対する一連のアクションをいかに学習するかで あり，本論文では，決定論的文圧縮と確率論的文圧縮の 2 つの手法を提案する． 提案手法は，主に，長文とそれらの圧縮文からなるコーパスにより評価された 機械学習法を使用する．どんな統計的機械学習法でも提案する文圧縮法に適用 可能ではあるが，本論文では，最大エントロピーモデルとサポートベクトルマシンモデルを用いる．評価実験により，提案する文圧縮法は従来法と比較して 良好な性能を得ることを示す．提案手法は従来法よりも人間が行う手法に近い 手法で文圧縮を行うというという特徴を持つ．さらに，提案する確率的文圧縮法は，文法や要旨の意味を保つという点に関して，決定論的文圧縮法を改善可能なことを示す．

異なる言語への機械翻訳に関しては，実例に基づく機械翻訳法の一種である翻 訳テンプレート学習に焦点を当てる．この手法には 2 つの欠点があり，それらは 学習フェーズと翻訳フェーズにある．学習フェーズでは，言語知識の不足によ り，翻訳テンプレート学習で使用されるテンプレートルールは膨大になり，翻訳に間違いが生じる原因になる．この問題を解決するため，新しい翻訳テンプ レート学習法を提案する．提案手法は shallow 構文解析を用いており，言語情報 をテンプレートルールに取り入れることが可能である．翻訳フェーズでのこの 手法の利点は，構文解析や意味解析などの複雑な解析を必要とせず，ルールに基づく機械翻訳の欠点を克服するということである．しかしながら，入力文と多数のテンプレートをマッチングする必要があり，翻訳結果が信頼性のないも のになり得るという欠点がある．さらに，テンプレートルールの多くは冗長であるにも関わらず，入力文に対して全てのルールを評価する必要があり，入力文が長い場合やテンプレートルールが多い場合は指数的な計算時間がかかるという問題がある．これらの点を考慮して，隠れマルコフモデル（hidden markov model: HMM）に基づく新しい手法を提案する．提案手法はルール集合に対して 制約を設定することで，冗長なルールの生成を抑制する．この手法はどんな 2 つ の言語対に対しても適用可能である．英語とベトナム語に対するアプリケーシ ョンを作成し，提案手法の翻訳精度と計算時間が従来法よりも優れていることを示す．

本論文では，重要文抽出と文圧縮で使用する学習データを，テキスト文書と要 約から自動的に作成する新しいアルゴリズムも提案する．最後に，アプリケー ション例として，英語のテキスト文書の要約をベトナム語に翻訳する CLTS シス テムを構築する．