

平成28年度 情報工学コース卒業研究報告要旨

外山 研究室	氏 名	佐 藤 充 晃
卒業研究題目	統計的機械翻訳の利用による法律文の自動要約システム	
<p>法律とは国会が制定する法の一形式であり、社会的規範のひとつである。法律文は一般の人々にとって読みにくいものであり、その内容を理解するのは容易ではない。官報では、公布された法律について、一般の理解に資するため「法令のあらまし」(以下、あらまし)が掲載されている。あらましは、公布された条約・法律・政令の公式な要約文書と言え、法律に関する情報の概要を理解する上で有用である。</p> <p>しかし、あらましは全ての法律に対して存在するわけではない。あらましは昭和48年4月から官報に掲載されているが、それ以前に公布された法律に対応するあらましは存在しない。実際、平成28年末時点で有効な法律1,966本のうち、官報にあらましが掲載される以前に公布された法律は750本もあり、これらの法律に対応するあらましは存在しない。これらの法律に対して、自動要約を施し、あらましに相当する要約文書を生成することができれば、法律の理解の促進に寄与するものと期待できる。</p> <p>一方、テキスト自動要約の分野では、統計的機械翻訳と同様に雑音のある通信路モデルにより自動要約の問題を捉えることができることに着目し、統計的機械翻訳を用いて文を自動要約する手法が提案されており、その有効性が示されている。</p> <p>そこで本研究では、法律からあらましに相当する文書を自動生成することを目指して、統計的機械翻訳を利用した法律文の自動要約システムを開発する。すなわち、法律文を統計的機械翻訳における原言語文、あらまし文を目的言語文とみなす。これらの間を対応付けたデータから統計的モデルを学習することにより、自動要約システムを構築する。</p> <p>本研究では、まず、統計的モデルの学習データを作成するため、法律とあらましの文書対を入力として、法律内の各文とあらまし内の各文との間の対応を自動抽出するアルゴリズムを開発した。あらましには、各文が法律のどの規定に関係した内容であることを示すため、「(第一条関係)」など関係規定を括弧で囲んで示した表現が記載されている。本アルゴリズムでは、この関係規定表現の括弧内の文字列と、その出現位置とを手がかりに、対応関係にある法律文とあらまし文の文ペアを抽出する。昭和50年から平成24年までに公布された新規制定法律1,150本とそれら各々に対応するあらましに対して本アルゴリズムを適用し、学習コーパス(3,190文ペア)を作成した。</p> <p>次に、作成した学習コーパスを用いて法律文の自動要約システムを構築した。法律文からあらまし文への翻訳モデルは学習コーパスの文ペアを、言語モデルは学習コーパス中のあらまし文を用いて学習した。また、デコーダには Moses を、翻訳モデル作成時における単語アライメントには GIZA++ を、言語モデルの学習には SRILM を、形態素解析には MeCab をそれぞれ利用した。</p> <p>本システムを評価するため、法律文に対する自動要約実験を実施した。テストデータには、昭和48年4月から昭和49年12月までに公布された新規制定法律58本とそれら各々に対応するあらましに対して、上述のアルゴリズムを適用して抽出した105文ペアを用いた。要約に対する自動評価指標である ROUGE により、本システムの出力文を評価した結果、ROUGE-1は78.03、ROUGE-2は70.31と高いスコアを示した。また、本システムの出力結果からランダムに取り出した20文について、正しく要約されているかを5段階で1名が主観的に評価した結果、平均3.2となった。これらの結果から、本システムがあらまし相当の要約文を生成できることが確認できた。</p>		