

Title	法令文書を対象とした並列構造解析の精緻化
Author(s)	松山, 宏樹
Citation	
Issue Date	2012-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10446
Rights	
Description	Supervisor: 白井清昭准教授, 情報科学研究科, 修士

法令文書を対象とした並列構造解析の精緻化

松山 宏樹 (1010062)

北陸先端科学技術大学院大学 情報科学研究科

2012 年 2 月 6 日

キーワード: 法令工学, 並列構造, 自然言語解析.

自然言語処理における並列構造解析は、解決が困難な問題の一つである。その原因として、文書ドメインごとに言語的性質が異なることが挙げられる。黒橋・長尾の研究では、並列関係にある句は互いに類似していると仮定し、句の類似度を DP マッチングで計算し、並列構造を解析している。これは実際に構文解析ツール KNP として実装されている。しかし KNP で法令文書を解析すると、特に並列構造解析の部分で解析誤りが頻出する。これは、法令文書は他の文書とは異なる言語的性質を有しているためと考えることができる。本論文では、法令文書を対象とした並列構造の検出及びその範囲の同定処理の精度向上を目的とし、法令文書に特化した並列構造を解析するための新しい手法を提案する。

本研究では並列構造を、1 つ以上の前方並列句、並列キー、後方並列句から構成されているとする。並列キーとは、並列構造で句を接続する働きをする語であり、前方及び後方並列句とは、並列キーの前方及び後方にある句で並列関係にあるものである。処理の流れを以下に述べる。まず並列キーを検出する。本研究では、「又は」「及び」「若しくは」「並びに」「と」「や」「かつ」「その他」の 8 つを並列キーとする。次に前方並列句の主辞を検出する。これは基本的には並列キーの直前の語である。次に後方並列句の候補を検出する。後方並列句の始点は基本的には並列キーの直後の語とする。後方並列句の終点は並列キーより後方にある語とし、前方並列句の主辞の品詞が助詞の場合は終点も助詞とし、主辞の品詞が動詞の場合は終点も動詞とする。名詞が主辞の場合、終点の候補は文節の最後の自立語で、かつ先に検出された前方並列句の主辞と類似度の高い上位 3 個の語とする。但し、読点、句点、他の並列キーに到達した時点で後方並列句の終点の探索を終了する。次に前方並列句の候補を検出する。前方並列句の終点は基本的には先に検出した主辞とし、始点は並列キーより前方にある文節の最初の単語とする。但し、読点、文頭、他の並列キーに到達した時点で前方並列句の始点の探索を終了する。次に得られた後方及び前方並列句の候補の類似度を計算する。並列関係にある句同士は互いに意味的に類似していると仮定し、全ての句の組み合わせから類似度の最も高い句の組を選択し、それぞれ前方並列句、後方並列句とする。句の類似度は、単語単位でアライメントをとり、対応関係にある単語の意味的類似度に基づいて算出する。また、対応関係のない単語があるときは句の

類似度も低くし、また、その単語が句の主辞に近い位置にあるほどそのペナルティを大きくする。更に、「第」「条」「項」「号」は法令文の条件番号を表わすのに使われる特別な語であることから、これらの語は同じ語に対応付けられるときのみ句の類似度を大きくした。次に既に決定された前方並列句の前に、別の前方並列句があるかをチェックし、ある場合はその候補を検出する。得られた前方並列句の候補と、すでに同定された前方及び後方並列句の類似度を計算し、それが最も高い候補を次の前方並列句とする。別の前方並列句が発見できなくなるまでこの処理を繰り返す。

次に、階層的な並列構造を解析する手法について述べる。本研究では、下位の並列構造を上位の並列構造解析よりも先に解析する必要があると考え、並列構造を構成する並列キーを下位の並列構造を構成するものから検出するようにした。すなわち、下位から上位の順に、ボトムアップ式に並列構造を解析する。解析を行う順序は、「又は」「及び」で結ばれる並列構造、「若しくは」「並びに」で結ばれる並列構造、「と」「や」「かつ」「その他」で結ばれる並列構造の順とした。これは、法令文では「又は」「及び」は内側の、「若しくは」「並びに」は外側の並列関係を記述するというルールを考慮したためである。また、階層的な並列構造解析を行う際に、前方並列句、後方並列句のどちらか一方に下位の並列構造が含まれるとき、両者の長さが大きく異なり、句の類似度を正確に見積もることができないという問題点がある。そこで、上位の並列構造の前方並列句、後方並列句が下位の並列構造を含むときは、それを下位の並列構造の後方並列句のみに置き換えることで、上位の並列句同士の長さのバランスをとるように工夫した。

提案手法に基づき、3つ以上の並列句を持つ並列構造や階層的な並列構造を解析するシステムを作成した。このシステムを用いて300文からなる法令文の解析を行い、検出された並列構造を評価した。その結果、並列構造検出のF値は50%、並列キー検出のF値は93%、前方並列句の検出のF値は65%、後方並列句の検出のF値は64%となった。また、提案手法をKNPと比較して評価したところ、提案手法はKNPよりも高い評価値を得た。KNPの並列構造のF値は26%であったのに対し、本研究では50%であり、KNPを24%上回った。