様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　１１　月　３０　日現在

機関番号：　１７１０２
研究種目：　挑戦的萌芽研究
研究期間：　2012～2015
課題番号：　２４６５００６５

研究課題名（和文）教師なし手法に基づく多言語形態素解析に関する研究

研究課題名（英文）Unsupervised Segmentation and Annotation of Texts

研究代表者
石井　久美子（田中久美子）（Tanaka-Ishii, Kumiko）

九州大学・システム情報科学研究科（研究院・教授

研究者番号：１０３２３５２８

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：与えられた文書に対して、目的に応じて文書内の単語単位を同定し、その範疇を定める処理は言語処理の要素技術である。たとえば、形態素解析の他、文書内の異言語部分を同定する処理などがある。本研究では、境界・範疇を同定する教師無し手法を考案する。三つの成果が得られた。第一に、文書内に埋め込まれた異言語部分を圧縮を用いて判定する方法を考案し、実用レベルにあることを示した。第二に、編集距離をBayes手法により拡張し、同じ意味を表す部分に対訳corpusを切り分ける方法を工夫した。第三に、最小オートマトンを用いて文のパターンを解析する手法を考案し、単位の切れ目と範疇を同定するための大規模な検証を行った。

研究成果の概要（英文）：This project aims at construction of unsupervized methods for automatic segmentation/annotation of given texts, a fundamental procedure of natural language processing. In addition to lemmatization, other tasks requring segmentation/annotation are also considered. Three achievements are obtained. First, using compression, we constructed an algorithm for detecting text subparts in other languages than the main text. Through a large scale experiment, the method was shown to work with a high accuracy applicable to text preprocessing. Second, the edit distance procedure was extended by Bayes method, and was applied to aligned corpora, to obtain translation pairs. Third, by use of minimal automaton, the patterns underlying sentences are detected, which serves for defining the segments within the sentence and further grouping of similarly used text parts.

研究分野：　Natural Language Processing

キーワード：　自然言語処理　形態素解析　教師無し学習　圧縮　Bayes手法

１．研究開始当初の背景

本研究で形態素解析とは、文を言語単位(たとえば単語など)に分解し、その単位の範疇(品詞など)を出力する解析を言う。検索・自動翻訳などを研究対象とする自然言語処理においては、このように単位の境界を得て、単位の範疇を定める処理は基礎となる要素技術である。特に、日本語は単語の分かち書きをしないため、単語への分解を行うことから処理は始まる。この処理は、日本語に限らず、分かち書きをしない言語では必要となる。

現在日本語には単語と品詞を同定する高性能な形態素解析器があるが、これは教師有り機械学習に基づき、新聞データを学習することにより構築されている。このため、新聞に見られる標準的な日本語に対する解析性能は極めて高い。一方で、昨今の言語処理が対象とする文面は、新聞だけではなく、SNSやブログなど多岐にわたり、文が短く簡易な言い回しが用いられているため既存の形態素解析による解析では限界があることも多い。また、グローバル化の現在、さまざまな言語の形態素解析が求められている。

形態素解析は、単語に分解して品詞を同定する以外にも、別の言語単位の境界を検出し、その単位の範疇を定める処理が発展として考えられ、さまざまに必要となる。たとえば、昨今では異なる言語の文書が混交することが多いが、その場合には言語処理の前処理として、言語ごとに文書を切り分け、言語を判定しなければならない。

２．研究の目的

本研究の目的は、教師無し機械学習手法に基づく形態素解析手法を考案することにある。形態素解析の概念を本研究では広く捉えており、与えられた文書に対して、処理目的に応じた単位境界とその範疇を定めることである。

教師無し機械学習手法は、解析の対象とする文面だけから解析を行い、別に学習のための教師データを必要としない。言い換えると、解析対象とする文書に最適な切れ目の同定を行い、類似の文脈下にある単位群を同じ範疇としてまとめる。このため、教師無し手法による形態素解析は、さまざまな多言語の文書に適用可能となるばかりか、通常の意味での形態素解析にとどまらず、境界を得て範疇を同定する処理一般に適用することが可能となる。

３．研究の方法

研究方法は、代表者が過去に提案した、単語境界を検出する手法を基礎とし、関連研究手法を参考に工夫し発展させることによる。代表者の過去の提案手法は、教師無し手法に基づくもので、単語境界において複雑さが増大するという性質がデータ内にあることを利用し、単語境界候補を得る。さらに文書全体で MDL(最小記述長)原理を用いて候補を絞り込み、境界を定める。この手法を基礎として、以下に述べる二つの既存研究動向を参考にし、独自手法を考案する。

第一に、本研究に関連する既存研究の中でも特に教師無し学習手法を吟味する。主として文の単語境界を得るための手法、ならびに、与えられる言語単位（単語など）の範疇をクラスタリングにより求めるものがある。いずれも、Bayes 手法に基づくものが主となる。

第二に、コーパス言語学と認知科学の最近の知見を参考にする。範疇や単語の境界とは何かを改めて問い直し、そこでの成果を工学的に応用する。

４．研究成果

成果としては以下の 4 つがある。

4-1.グローバル化の現在、文書には複数の言語が混じっている。文書はある言語で書かれているが、その中に異なる言語部分が埋め込まれている。これを同定し、部分の言語も判定することが、文書処理の前処理として必要となる。本研究では MDL 原理に動的計画法を組み合わせ、独自のアルゴリズムを考案した。３００言語を超える言語間で大規模な実験を行い、異言語部分判定を実用レベルに達する性能で実現できることを示した。成果は自然言語分野最難関の国際会議 ACL にて発表した。

4-2.　文の様相を判定する部分を切り出し、文の様相の自動判定性能を向上させる方法を考案した。文内で文意に関わる特殊部分を自動で切り出して範疇化する点で、本研究の成果の一部として位置付けることができる。成果は国内の雑誌論文として採録となった。

4-3. 文字列の対応をとるアルゴリズムに編集距離があるが、これを既存の Bayes 手法を参考に教師無し手法として発展させた。これを用いると、対訳関係にある文書(以下対訳コーパス)に対して、教師無し手法で対訳関係にある文書部分を同定する(すなわち、意味の同一性に基づいて境界を定める）ことができる。Bayes 編集距離を用いて、対訳コーパスを二言語同時に形態素解析し、翻訳関係にある部分を獲得することを行った。

国内学会での２編の発表を行ったのち（査読無)、現在英文論文誌に２編投稿中である(査読有)。１編は現在条件付採録である。もう１編は機械学習分野で最上位の論文誌に投稿中で、1 年半以上たって未だ査読結果を待っている状況である。

本研究は NICT との共同研究を通して行った。

4-4. オートマトンを利用して、文のパターン解析を行う手法を考案した。本研究は、コーパス言語学の最新の成果からヒントを得たもので、それによると文法とは本来的にパターンであり、パターンに境界、範疇を不可分に含むという。実際、形態素解析が文構造解析と不可分であることは、自然言語処理分野でも言われている。本研究では構造を教師無し手法で解析することから、境界や範疇を得る。与えられた文字列集合を受理するオートマトンの中でも、節点数が最小となる最小オートマトンがあるが、本研究では最小オートマトンを用いる。共通の構造を持つ文集合に対して最小オートマトンと等価な構造を作成し、後処理を行うと、そこに、共通の文構造が穴空きのパターンとして表現される。たとえば、"regard A as B"のパターンを含む複数の文からこの構造が現れる：A、B の前後で境界があること、また A,B に含まれるものは同範疇として同定することができる。

得られる構造の評価を大規模に検証した。現在、論文発表を準備中である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2　件）

- Andrew Finch, Taisuke Harada, Kumiko Tanaka-Ishii, and Eiichiro Sumita. Inducing a bilingual lexicon from short parallel multiword sequences. Transaction on Asian Low-resource Language Information Processiong, 2016. 採録 to appear.

- Andrew Finch, Nakatani Koki, Kumiko Tanaka-Ishii, and Eiichiro Sumita. Stochastic block edit distance. 投稿中

- Daiki Hirano, Kumiko Tanaka-Ishii and Andrew Finch. Pattern Extraction using Minimal Automaton. 2016, 投稿中.

- Andre Horie and Kumiko Tanaka-Ishii. Sentence hedge detection without cue annotation: A heuristic cue selection approach. 自然言語処理, 21(1):27-40, 2014.

〔学会発表〕（計　4　件）

- Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. Text segmentation by language using minimum description length. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers pages 969-978, July 2012.

- 中谷洸樹, Andrew Finch, 田中久美子,

and 隅田英一郎. 確率的ブロック編集距離. 言語処理学会大会論文集, pages 352--355, 2014.

- 原田泰佑, Andrew Finch, 田中久美子, and 隅田英一郎. Transliteration の拡張としての Wikipedia からの意味的翻訳対の抽出. 言語処理学会大会論文集, pages 417--420, 2015.

〔図書〕(計　1　件)

- Kumiko Tanaka-Ishii. Consonants as skeleton of language: Statistical evidences through text production. In Language Production, Cognition, and the Lexicon, Chapter 16, pages 287-297. Springer, 2014.

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者（本人）
田中久美子
(Tanaka-Ishii, Kumiko)
九州大学・大学院システム情報科学研究院・教授　）
研究者番号：10323528

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：