科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１２　日現在

機関番号：　１４３０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００１７７

研究課題名（和文）グラフ理論に基づく自然言語解析の定式化

研究課題名（英文）Formulation of Natural Language Processing　based on Graph Theory

研究代表者
森　信介（Mori, Shinsuke）

京都大学・学術情報メディアセンター・准教授

研究者番号：９０４５６７７３

交付決定額（研究期間全体）：（直接経費）　　4,100,000円、（間接経費）　　1,230,000円

研究成果の概要（和文）：係り受け解析において、一部の単語にのみ情報が付与された部分的アノテーションコーパスの利用とそれを用いることで最高水準の精度と高い分野適応性を実現した。研究代表者および連携研究者は、係り受け解析の最小全域木問題としての定式化(文献[3])を参考に実装を行った。このようにして実現された言語処理システムをツールとして公開した。また、学習のためのコーパスについても学会発表を行った上で、公開した。

研究成果の概要（英文）：We realized a dependency parser trainable from partially annotated corpus in which only a few words are annotated with dependency information. The parser works as accurate as state-of-the-art parsers. In detail, the parser is based on a maximum spanning tree framework. We made the natural language processing tools publicly available. In addition, we made a presentation about the corpus for training and made it public.

研究分野：　知能情報

科研費の分科・細目：　自然言語処理

キーワード：　係り受け解析　部分的アノテーション

１．研究開始当初の背景
自動単語分割と係り受け解析は、新聞などの学習データがある特定の分野において高い解析精度が実現されており、その成果もツールとして利用可能である。しかし、実際には、自然言語処理の要求のほとんどは、ツールが想定していな分野のテキストが対象であり、解析精度の著しい低下が大きな問題となっている。この問題に対してユーザーが取り得る手段は、適応分野特有の単語を辞書に追加する程度であり、全く不十分である。

自動単語分割におけるこの問題に対して、研究代表者は、一部の単語にのみ情報が付与された部分的アノテーションコーパスの利用を提唱し、それを利用可能な自動単語分割器を設計し、世界最高水準の精度と高い分野適応性を実験的に示すとともにツールとして公開した(文献[1,2])。本研究計画では、係り受け解析においても、世界最高水準の精度と高い分野適応性を実現し、係り受け解析の応用を促進する。研究代表者および連携研究者はすでに、係り受け解析の f 最小全域木問題としての定式化(文献[3])を参考に簡単な実装を行った。その結果、部分的アノテーションコーパスにより高い分野適応性を実現する見通しを得ている。

使役・受動態などの格変換や照応・省略の補完は、現在一部研究があるものの、入出力を含めた問題の定義(仕様)について研究者間での共有認識がない。そのため、他分野の研究者や言語処理システムの開発者が容易に用いることができるツールは存在しない。
２．研究の目的
本研究計画では、まず、これらの問題を単語をノードとするグラフの問題と考え、グラフ理論を用いて定式化する。次に、部分的アノテーションコーパスにより高い分野適応性を実現する。これら、高次の言語処理の実現には、部分的アノテーションコーパスは必須である。すなわち、文全体に情報付与する現在の枠組みでは、照応などの高次の言語現象のデータを機械学習に十分な量とするためには、単語分割などの低次の言語現象に対する大量の情報付与が必要となる。これに対して、代表者が提案する部分的アノテーションの枠組みでは、必要となる情報付与量が圧倒的に少量で済み、様々な分野における高い精度の言語処理が実現可能となる。最後に、このようにして実現された言語処理システムをツールとして公開する。
３．研究の方法
本研究では、グラフ理論に基づく統一的な言語処理の提案と実装を目指し、以下の点を明らかにする。

単語単位の係り受け解析のグラフ理論的定式化:主に日本語を対象として、係り受け解析を単語をノードとするグラフの最小全域木問題としての定式化(文献[3])を拡張し、一部の単語にのみ係り受け情報を付与したデータから学習する枠組みを実現する。

受動態と能動態などの格変換をグラフ変換として定式化:単語単位の係り受け木において、動詞と主語・目的語の関係は、部分木として表現される。受動態と能動態の変換を部分木への変換操作として定式化する。同様の定式化を使役や連体修飾にも拡張し、これを実装・公開する。

照応・省略の補完の定式化:照応・省略などの言語現象もグラフ理論を用いて定式化可能であることを示す。例えば、ゼロ代名詞の補完は、先行詞となる名詞のノードを適切な格助詞をラベルとするノードを介して動詞ノードに接続するシュタイナー木(追加可能なノード集合が与えられた最小全域木)として定式化できる。このような定式化を一般的な照応・省略に拡張する。

４．研究成果
係り受け解析において、一部の単語にのみ情報が付与された部分的アノテーションコーパスの利用とそれを用いることで最高水準の精度と高い分野適応性を実現した。研究代表者および連携研究者は、係り受け解析の最小全域木問題としての定式化(文献[3])を参考に実装を行った。このようにして実現された言語処理システムをツールとして公開した。また、学習のためのコーパスについても学会発表を行った上で、公開した。
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 11 件）全て査読有り
http://plata.ar.media.kyoto-u.ac.jp/mori/research/public/

Substring-based Machine Translation
Graham Neubig, Taro Watanabe, Shinsuke Mori, Tatsuya Kawahara
Machine Translation, Volume 27, Issue 2, pp.139-166, 2013.

A Pointwise Approach to Training Dependency Parsers from Partially Annotated Corpora
Daniel Flannery, Yusuke Miyao, Graham Neubig, Shinsuke Mori
Natural Language Processing, Vol.19, No.3, pp.167-191, September, 2012.

自然言語処理における分野適応
森 信介
人工知能学会誌, Vol.27, No.4, pp.365-372, 2012.

A Monotonic Statistical Machine Translation Approach to Speaking Style Transformation
Graham Neubig, Yuya Akita, Shinsuke Mori, Tatsuya Kawahara
Computer Speech and Language, Vol.26, Iss.5, pp.349-370, October 2012.

Joint Phrase Alignment and Extraction for Statistical Machine Translation
Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, Tatsuya Kawahara
Journal of Information Processing, 2012.

Bayesian Learning of a Language Model from Continuous Speech
Graham Neubig, Masato Mimura, Shinsuke Mori, Tatsuya Kawahara
IEICE, Vol.E95-D, No.2, pp.614-625, 2012.

音声対話システムにおける簡略表現認識のための自動語彙拡張
森 信介, 駒谷 和範, 勝丸 真樹, 尾形 哲也, 奥乃 博
情報処理学会論文誌, Vol.52, No.12, pp.1882-7764, 2011.

述語項の類似度に基づく情報抽出・推薦を行う音声対話システム
吉野 幸一郎, 森 信介, 河原 達也
情報処理学会論文誌, Vol.52, No.12, pp.1882-7764, 2011.

点予測による単語分割
森 信介, Neubig Graham, **坪井 祐太**
情報処理学会論文誌, Vol.52, No.10, pp.2944-2952, 2011.

点予測による形態素解析
森 信介, 中田 陽介, Neubig Graham, **河原達也**
自然言語処理, Vol.18, No.4, pp.367-381, 2011.

確率的タグ付与コーパスからの言語モデル構築
森 信介, 笹田 鉄郎, Neubig Graham
自然言語処理, Vol.18, No.2, pp.71-87, 2011.

〔学会発表〕（計 20 件）全て査読有り
http://plata.ar.media.kyoto-u.ac.jp/mori/research/public/

A Japanese Word Dependency Corpus
Shinsuke Mori, Hideki Ogura, Tetsuro Sasada
LREC, 2014.

Language Resource Addition: Dictionary or Corpus?
Shinsuke Mori, Graham Neubig
LREC, 2014.

Flow Graph Corpus from Recipe Texts
Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, Tetsuro Sasada
LREC, 2014.

Feature Extraction and Summarization of Recipes using Flow Graph
Yoko Yamakata, Shinji Imahori, Yuichi Sugiyama, Shinsuke Mori, Katsumi Tanaka
SocInfo, 2013.

Predicate Argument Structure Analysis using Partially Annotated Corpora
Koichiro Yoshino, Shinsuke Mori, Tatsuya Kawahara
IJCNLP, pp.957-961, 2013.

A Framework and Tool for Collaborative Extraction of Reliable Information
Graham Neubig, Masahiro Mizukami, Shinsuke Mori
Workshop on Language Processing and Crisis Information (LPCI), 2013.

Incorporating Semantic Information to Selection of Web Texts for Language Model of Spoken Dialogue System
Koichiro Yoshino, Shinsuke Mori, Tatsuya Kawahara
ICASSP, pp.8252-8256, 2013.

Language Modeling for Spoken Dialogue System based on Filtering using Predicate-Argument Structures
Koichiro Yoshino, Shinsuke Mori, Tatsuya Kawahara
Coling, 2012.

Statistical Input Method based on a Phrase Class n-gram Model
Hirokuni Maeta, Shinsuke Mori
WTIM, pp.1-13, 2012.

An Ensemble Model of Word-based and Character-based Models for Japanese and Chinese Input Method
Yoh Okuno, Shinsuke Mori
WTIM, pp.15-27, 2012.

IwaCam: a Multimedia Processing Platform for Supporting Video-Based Cooking Communication
Hidenori Tsuji, Yoko Yamakata, Takuya Funatomi, Hiromi Hiramatsu, Shinsuke

Mori
FGCT (Int'l Conf. on Future Generation Communication Technology), 2012.

Language Modeling for Spoken Dialogue System based on Sentence Transformation and Filtering using Predicate-Argument Structures
Koichiro Yoshino, Shinsuke Mori, Tatsuya Kawahara
APSIPA, 2012.

A Machine Learning Approach to Recipe Text Processing
Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, Koichiro Yoshino
Cooking with Computers Workshop, August, 2012.

Inducing a Discriminative Parser to Optimize Machine Translation Reordering
Graham Neubig, Taro Watanabe, Shinsuke Mori
EMNLP-CoNLL, 2012.

Machine Translation without Words through Substring Alignment
Graham Neubig, Taro Watanabe, Shinsuke Mori, Tatsuya Kawahara
ACL, 2012.

Training Dependency Parsers from Partially Annotated Corpora
Daniel Flannery, Yusuke Miyao, Graham Neubig, Shinsuke Mori
IJCNLP, pp.776-784, 11/10, 2011.

Discriminative Method for Japanese Kana-Kanji Input Method
Hiroyuki Tokunaga, Daisuke Okanohara, Shinsuke Mori
WTIM, 2011

A Pointwise Approach to Pronunciation Estimation for a TTS Front-end
Shinsuke Mori, Graham Neubig
InterSpeech, 08/28, 2011.

An Unsupervised Model for Joint Phrase Alignment and Extraction
Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, Tatsuya Kawahara
ACL-HLT, 2011.

Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis
Graham Neubig, Yosuke Nakata, Shinsuke Mori
ACL-HLT, 2011.

〔図書〕（計 0 件）

〔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 0 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
http://plata.ar.media.kyoto-u.ac.jp/mori/research/

６．研究組織
(1)研究代表者
森 信介（MORI Shinsuke）
京都大学学術情報メディアセンター准教授
研究者番号：90456773

(2)研究分担者
（ ）

研究者番号：

(3)連携研究者
（ ）

研究者番号：