様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 5 月 31 日現在

機関番号：14301
研究種目：若手研究（A）
研究期間： 2008～2010
課題番号：20680008
研究課題名（和文） 係り受けや照応・省略などの高次言語情報を用いた確率的言語モデル

研究課題名（英文） Stochastic language modeling using high linguistic information

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

研究成果の概要（和文）：まず、点予測による手法を提案し、自動単語分割の精度向上を実現した。つぎに、単語境界情報と係り受け情報が付与されたコーパスを辞書の例文と経済新聞記事から作成した。また、確率的なアノテーションを提案し、確率的単語分割コーパスや確率的読み付与コーパスからの言語モデルの作成を提案した。

研究成果の概要（英文）：First we proposed a pointwise method and realized an improvement of word segmentation. Then we created a corpus consisting of dictionary example sentences and newspaper articles annotated with dependency information. We also proposed stochastic annotation and language model building from a stochastically segmented or tagged corpus.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
20 年度	6,100,000	1,830,000	7,930,000
21 年度	3,100,000	930,000	4,030,000
22 年度	3,100,000	930,000	4,030,000
年度			
年度			
総 計	12,300,000	3,690,000	15,990,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：係り受け, 照応・省略, 確率的言語モデル, 認知科学, 音声認識

１．研究開始当初の背景

言語処理研究は、対象とする言語現象に応じて、様々な段階に分割されている。まず形態素解析を行ない、その出力を構文解析し、格解析を行なって、その出力を談話解析している。後段の処理によって得られる情報の参照(例えば照応詞の指示対象を考慮した係り受け先の決定)が長らく未解決の問題となっているが、この解決に正面から取り組む研究は少ない。

２．研究の目的

本研究計画では、係り受けや照応・省略などの高次の言語現象を記述する統一的モデルについて考察し、文字列から照応・省略までを統一的に扱うことで、上述の問題を解決する。

３．研究の方法

まず、言語資源に、単語境界情報と単語間の係り受け情報を付与し、分析・実験のためのコーパスを作成する。次に、文字列から係り受けまでの言語現象を対象とする構造的言語モデルを構築し、構文を考慮した音声認識や仮名漢字変換システム、あるいは文字列を入力とし単語分割と構文解析を同時に行

なうシステムを作成し、既存手法との比較などの評価実験を行なう。

次に、構造的言語モデルをより高度な言語現象(照応や省略)や話し言葉(倒置や誤り)を記述するモデルに発展させる。このモデルを用いることで、文字列から照応・省略までを統一的に扱い、単語分割と構文解析と照応解析を同時に行なうことが可能になる。これにより、後段の処理によって得られる情報のフィードバックという言語処理の長年の課題の一つの解決を与える。

４．研究成果

まず、点予測による手法を提案し、自動単語分割の精度向上を実現した。実験では、単語と品詞として国立国語研究所が提案する定義を採用し、その基準に沿う『現代日本語書き言葉均衡コーパス』を用いて有効性を示した。特に、部分的アノテーションコーパスの概念を提案し、これを含むさまざまな言語資源からの学習が可能であることを示し、安価かつ高速の分野適応が実現した。点予測による手法を品詞推定や読み推定にも適用し、同様の有効性が実現できることを示した。この成果は、「言語処理ソフトウェア KyTea」として公開しており、多数に利用されている。また、分野適応のための能動学習ツールを構築し、公開している。これらは、東北大震災の安否情報の言語処理に使われた。

高精度の自動係り受け解析を実現するために、単語境界情報と係り受け情報が付与されたコーパスを辞書の例文と経済新聞記事から作成した。それぞれ、13,000 文と 10,025 文からなる。点予測による係り受け解析を実現し、これらのコーパスを用いて既存手法と同等の精度が得られることを示した。また、係り受け情報が付与されたコーパスから、構造情報を利用する確率的言語モデルを構築し、予測力における有効性を確認した。

また、確率的なアノテーションを提案し、確率的単語分割コーパスや確率的読み付与コーパスからの言語モデルの作成を提案し、新聞やウェブのデータから大規模な言語モデルを作成し、音声認識や仮名漢字変換への応用を行った。仮名漢字変換については、エンジンをフリーのソフトウェアとして公開している.

o 言語処理ソフトウェア KyTea：
http://www.phontron.com/kytea/index-ja.html
o 仮名漢字変換ソフトウェア SIMPLE：

http://plata.ar.media.kyoto-u.ac.jp/mori/research/topics/KKC/

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 5 件）

1. 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，To appear in 自然言語処理，Vol.18，No.2，2011.（査読有り）

2. 3 種類の辞書による自動単語分割の精度向上，森 信介，小田 裕樹，To appear in 自然言語処理，Vol.18，No.2，2011.（査読有り）

3. 自動獲得した未知語の読み・文脈情報による仮名漢字変換，笹田 鉄郎，森 信介，河原 達也，自然言語処理，Vol.17，No.4，pp.131-154，2010.（査読有り）

4. 擬似確率的単語分割コーパスによる言語モデルの改良，森 信介，小田 裕樹，自然言語処理，Vol.16，No.5，pp.7-22，2009.（査読有り）

5. 日本語単語分割の分野適応のための部分的アノテーションを用いた条件付確率場の学習，坪井 祐太，森 信介，鹿島 久嗣，小田 裕樹，松本 裕治，情報処理学会論文誌 Vol.50，No.6，pp.1622-1635，2009.（査読有り）

〔学会発表〕（計 15 件）

1. A Pointwise Approach to Pronunciation Estimation for a TTS Front-end, Shinsuke Mori, Graham Neubig, InterSpeech, Firenze, Italy, 29th Aug., 2011. (To appear)

2. An Unsupervised Model for Joint Phrase Alignment and Extraction, Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, Tatsuya Kawahara, ACL-HLT, Portland, USA, 20th June, 2011. (To appear)

3. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, Graham Neubig, Yosuke Nakata, Shinsuke Mori, Portland, USA, 20th June, ACL-HLT, 2011. (To appear)

4. 点予測と系列予測の 2 段階化による品詞推定の精度向上，中田 陽介，NEUBIG Graham，森 信介，河原 達也，情報処理学会研究報告，NL200，東京，1 月 28 日，2011.

5. 点予測による形態素解析，中田 陽介，NEUBIG Graham，森 信介，河原 達也，情報処理学会研究報告，NL-198，東京，9 月 17 日，

2010.

6. 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，情報処理学会自然言語処理研究会，NL-196/SLP81，東京，5 月 27 日，2010.

7. Word-based Partial Annotation for Efficient Corpus Construction, Graham Neubig, Shinsuke Mori, LREC, Valetta, Malta , 20th May, 2010.

8. 点推定と能動学習を用いた自動単語分割器の分野適応，Neubig Graham，中田 陽介，森 信介，言語処理学会第 16 回年次大会，東京，3 月 11 日, 2010.

9. 利用過程で得られる言語情報を活用する音声言語処理システム，森 信介，前田 浩邦, NLP 若手の会 第 4 回シンポジウム，京都, 10 月 1 日, 2009.

10. 3 種類の辞書による自動単語分割の精度向上，森 信介，小田 裕樹，情報処理学会自然言語処理研究会，NL-193，京都，9 月 29 日, 2009.

11. Automatic Word Segmentation using Three Types of Dictionaries, Shinsuke MORI, Hiroki ODA, PACLING, Sapporo, Japan, 1st Sep., 2009.

12. Extracting Word-Pronunciation Pairs from Comparable Set of Text and Speech, Tetsuro SASADA, Shinsuke MORI, Tatsuya KAWAHARA, InterSpeech, pp.1821-1824, Brisbane, Australia, 22nd Sep., 2008.

13. Training Conditional Random Fields Using Incomplete Annotations, Yuta TSUBOI, Hisashi KASHIMA, Shinsuke MORI, Hiroki ODA, Yuji MATSUMOTO, Coling, Manchester, UK , 18st Aug., 2008.

14. 音声認識のための言語処理：何が足りないか？，森 信介，情報処理学会音声言語情報処理研究会，SLP-72，盛岡，7 月 18 日，2008.

15. テキストと音声を用いた単語と読みの自動獲得，笹田 鉄郎，森 信介，河原 達也，情報処理学会音声言語情報処理研究会，SLP-72，盛岡，7 月 18 日，2008.

〔図書〕（計 1 件）
1. 言語処理学事典，2.1.1 n グラムモデル，2.1.3 言語モデルの評価，2.9.5 隠れマルコフモデル，言語処理学会編，森 信介，他多数，共立出版株式会社，2009.

〔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）

研究者番号：90456773

(2)研究分担者
無し

(3)連携研究者
無し