様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 6 月 7 日現在

研究種目：基盤研究（B）
研究期間： 2007 ～ 2009
課題番号： 19300049
研究課題名（和文）
非線形言語モデルによる日本語単文の意味解析基盤技術
研究課題名（英文）
Non-Compositional Language Model and Pattern Dictionary Development for Japanese Simple Sentences
研究代表者
村上仁一（MURAKAMI JIN'ICHI）
鳥取大学・工学研究科・准教授
研究者番号：90304196

研究成果の概要（和文）：

本研究では，まず単文の日英文パターン辞書を用いて日英パターン翻訳を行い， 文法構造を英語に近づける．この日英文パターンが持つ大局的な文法情報を用いることで N-gram モデルにおける局所的な構文問題が解消できると考えた．そして出力文に対し，統計翻訳でさらに英英翻訳を行う．この処理により，局所的な修正を行うことで翻訳精度が向上すると考えた.実験の結果,従来の日英統計翻訳システムと比べて提案手法のシステムでは，文パターンの文法情報が多く残されている場合に翻訳精度が高く，有効性が 確認出来た.

研究成果の概要（英文）：

We have developed a two-stage machine translation (MT) system. The first stage is a rule-based machine translation system. The second stage is a normal statistical machine translation system. For Japanese-English machine translation, first, we used a Japanese-English rule-based MT, and we obtained "ENGLISH" sentences from Japanese sentences. Second, we used a standard statistical machine translation. This means that we translated "ENGLISH" to English machine translation. We believe this method has two advantages. One is that there are fewer unknown words. The other is that it produces structured or grammatically correct sentences. From the results of experiments, our proposed method was effective for the Janpanese to English machine translation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2007 年度	4,300,000	1,290,000	5,590,000
2008 年度	3,400,000	1,020,000	4,420,000
2009 年度	2,000,000	780,000	3,380,000
年度			
年度			
総 計	10,300,000	3,090,000	13,390,000

研究分野：機械翻訳，音声認識
科研費の分科・細目：情報学・知能情報学
キーワード：単語モデル，非線形性，文型パターン，単文，機械翻訳，統計翻訳

１．研究開始当初の背景

結合価パターン方式では，助詞や助動詞を用いて表される主体的表現の部分は分離して処理しなければならないため，それらの持つ非線形な意味が失われるという問題がある．この問題を解決するために，CREST の研究費をもらって，重文複文を対象として，表現意味辞書を研究開発した．用例数が２５万文，パターン数は，３０万パターンに達する．また，最終的な被覆率は 80%となった．

２．研究の目的

CREST の研究費をもらって表現意味辞書を研究開発したが，この対象は重文複文であった．しかし，この技術は単文にも応用できる．そこで日本語単文を対象に，その内容（意味）を取り出し，変換，加工するための意味解析基盤技術を確立する．また，その応用例として，意味的等価変換方式に適用し効果を確認する．

（1）「文型パターン辞書」の研究開発
（2）文型パターン辞書の縮退化
（3）文型パターン辞書の意味類型化
（4）適合パターン選択方式の確立
（5）意味的等価変換方式への適用実験

３．研究の方法

１．文型パターン辞書の研究開発

日英単文の対訳コーパス 40 万件を対象に単文パターン化を行い，昨年度の結果と組み合わせて，単文パターン辞書を作成する．単文対訳こ一端からのパターン作成の手順はおおよそ以下の通りである．

（1）単文抽出と対訳標本の抽出
（2）単文文型パターン化方法の検討と設計
（3）汎化作業支援プログラムの作成
（4）単文対訳標本の汎化作業
（5）単文パターン辞書の評価と改良

３．文型パターン辞書の意味類型化

「意味類型化」は，文型パターンを意味によって分類することであるが，各文型パターンの意味を概念（真理項と呼ぶ）を用いて定義することができれば，文型パターンの意味分類は可能となる．そこで，本研究では，単文の意味分類体系を作成し，１で作成し，２で縮退化した「文型パターン辞書」の各パターンに対して，意味分類コードを付与する．

（1）意味類型化方式の設計
（2）文型パターン辞書の意味類型化
（3）意味検索プログラムの設計と試作
（4）意味類型パターンの改良

４．研究成果

作成した単文パターンを利用して日英翻訳システムを試作した．まず，本研究では日本語-英語間における大きく異なる文法構造に着目し,まずパターン翻訳を行うことで文法構造を 英語に近づけた．次に統計翻訳を行うことで文法構造に対し,局所的な修正を行う手法を提案した.実験の結果，日英文パターン辞書における変数の個数が少ない場合に は,翻訳精度が向上した.今後は，より精度の高いパターンの作成を試みる．

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 3 件）

①村上仁一，鏡味良太，徳久雅人，池原悟，統計翻訳における人手で作成された大規模フレーズテーブルの効果， 自然言語処理 2010 年 7 月号，2010，査読有

②徳久 雅人, 村上 仁一, 池原 悟， 漫画における表情に着目した情緒タグ付きテキスト対話コーパスの構築
自然言語処理, Vol.14, No.3, pp.192-217, 2007，査読有

③Satoru Ikehara, Masato Tokuhisa, Jin'ichi Murakami Masashi Saraki, Masahiro Miyazaki, Naoshi Ikeda, Pattern Dictionary Development based on Non-Compositional Language Model for Japanese Compound and Complex Sentences, International Journal of Computer Processing of Oriental Language, Vol.20, Nos.2 , 2 & 3 pp.151-163, 2007，査読有

〔学会発表〕（計 27 件）

①Masato Tokuhisa, Jin'ichi Murakami, Satoru Ikehara: Affective Blog Analyzer - What People feel to, Proceedings of the 2nd International Conference on Agents and Artificial Intelligence(ICAART2010), Vol.1, pp.247-252, 2010, 査読有

②吉田大蔵, 徳久雅人,村上仁一, 池原悟, 格助詞およびその相当表現のパターン翻訳の試み,電子情報通信学会技術研究報告, 思考と言語, TL2009-43, pp.13-18, 2010, 査読無

③東江 恵介, 村上 仁一, 徳久 雅人,池原 悟, 日英統計翻訳における英辞郎の効果, 言語処理学会第 16 回年次大会, PB2-3, pp.641-644, 2010, 査読無

④西村 拓哉, 村上 仁一, 徳久 雅人, 池原 悟, 文単位のパターンを用いた統計翻訳, 言語処理学会第 16 回年次大会, PB2-12, pp.676-679, 2010, 査読無

⑤福田智大, 村上 仁一, 徳久 雅人, 池原 悟, ルールベース翻訳を前処理に用いた統計翻訳, 言語処理学会第 16 回年次大会, PB2-12, pp.676-679, 2010, 査読無

⑥猪澤 雅史, 村上 仁一, 徳久 雅人, 池原 悟, 文節区切りの学習データを用いた, 日英統計翻訳の検討, 言語処理学会第 16 回年次大会, B5-7, pp.1022-1025, 2010, 査読無

⑦Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara, Statistical Machine Translation adding Pattern-based Machine Translation in Chinese-English Translation, International Workshop on Spoken Language Translation 2009, (IWSLT 2009), pp.107-112, October 2009, 査読有

⑧滝川晃司, 徳久雅人,村上仁一, 池原悟,情緒推定用パターン辞書における荒いレベルの情緒原因判断条件, 電子情報通信学会技術研究報告, 言語理解とコミュニケーション, NLC2009-40, pp.43-48, 2009, 査読無

⑨福田泰介, 徳久雅人,村上仁一, 池原悟 受動態に変換した結合価パターン辞書 電子情報通信学会技術研究報告, 思考と言語, TL2009-43, pp.19-24, 2009, 査読無

⑩村上 仁一, 徳久雅人, 池原 悟,パターン翻訳と統計翻訳の結合, 言語処理学会 2008 年度年次大会, pp.120-123, 2009, 査読無

⑪鏡味 良太, 村上 仁一, 徳久雅人, 池原 悟, 統計翻訳における人手で作成された大規模フレーズテーブルの効果,言語処理学会 2008 年度年次大会, pp.224-227, 2009, 査読無

⑫岡崎 弘樹, 村上 仁一, 徳久雅人, 池原 悟, 日本語文法構造の変換による日英統計翻訳, 言語処理学会 2008 年度年次大会,pp.240-243, 2009, 査読無

⑬大友 謙一, 村上 仁一, 徳久雅人, 池原 悟 WHY 型 QA システムにおける回答抽出方法の改良, 言語処理学会 2008 年度年次大会, pp.586-589, 2009, 査読無

⑭滝川 晃司, 徳久 雅人, 村上 仁一, 池原 悟, 情緒推定用パターン辞書における情緒原因判断条件の改良, 言語処理学会 2008 年度年次大会, pp.829-832, 2009, 査読無

⑮Satoru Ikehara, Masato Tokuhisa, Jin'ichi Murakami, Non-Compositional Language Model and Pattern Dictionary Development for Japanese Compound and Complex Sentences, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp.353-360, 2008, 査読有

⑯Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara , Statistical Machine Translation without Long Parallel Sentences for Training Data, International Workshop on Spoken Language Translation 2008, (IWSLT 2008), pp.132-137, 2008, 査読有

⑰Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara, Statistical Machine Translation with Long Phrase Table and without Long Parallel Sentences, NII Test Collection for IR Systems 07 WorkShop, (NTCIR07 2008), pp.454-461, 2008, 査読有

⑱猪澤 雅史, 村上仁一, 徳久雅人, 池原悟, 統計翻訳における, 単文と重文複文の翻訳精度の評価, 情報処理学会研究報告, 自然言語処理, 2008-NL-188, pp.79-84, 2008, 査読無

⑲中道龍三, 徳久雅人, 村上仁一, 池原悟, 情緒推定の手がかりとなる接続表現の収集, 電子情報通信学会技術研究報告, 思考と言語, TL2008-44, pp.1-6, 2008, 査読無

⑳前田浩佑, 徳久雅人, 村上仁一, 池原悟,情緒傾向値付きパターン辞書を用いた文末表現の分析,電子情報通信学会技術研究報告, 思考と言語, TL2008-47, pp.19-24, 2008, 査読無

(21)田村 元秀, 村上 仁一, 徳久 雅人, 池原 悟, Web 検索エンジンを用いたWhy型質問応答システムに関する研究, 情報処理学会研究報告, 自然言語処理, 2008-NL-183, pp.6-15 2008, 査読無

(22)楊 鵬, 村上 仁一, 徳久 雅人, 池原 悟, 結合価パターンを用いた日中機械翻訳システムの構築, 情報処理学会研究報告, 自然言語処理, 2008-NL-183, pp.121-126, 2008, 査読無

(23)徳久雅人, 前田浩佑, 村上仁一, 池原悟, 対話行為と情緒を解析するための文末表現パターンの作成, 電子情報通信学会技術研究報告, 言語理解とコミュニケーション, NLC2007-95, pp.45-50, 2008, 査読無

(24)Satoru Ikehara, Masato Tokuhisa, Jin'ichi Murakami, Analogical Mapping Method and Semantic Categorization of Japanese Compound and Complex Sentence Patterns, Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pp.181-190, 2007, 査読有

(25)Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara, Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables, International Workshop on Spoken Language Translation 2007, pp.151-155, 2007, 査読有

(26)水田 理夫, 村上 仁一, 重文・複文文型パターン辞書を用いた日英パターン翻訳システムにおける離散記号の解決手法, LACE第12回, 2007, 査読無

(27)徳久雅人, 前田浩佑, 村上仁一, 池原悟, 心的状態を表す対話行為タグ付きテキスト対話コーパスの構築, 電子情報通信学会技術研究報告, 思考と言語, TL2007-45, pp.25-30, 2007, 査読無

〔図書〕(計1件)
①池原悟, 岩波出版, 非線形言語モデルによる自然言語処理, 2009年, 324頁

〔その他〕
ホームページ等
http://unicorn.ike.tottori-u.ac.jp/paper

６．研究組織

(1)研究代表者
池原 悟 (IKEHARA SATORU)
鳥取大学・工学研究科・教授
研究者番号：70283968
(H19.3~H21.12)

(2)研究代表者
村上仁一 (MURAKAMI JIN'ICHI)
鳥取大学・工学研究科・准教授
研究者番号：90304196
(H21.12~H22.03)

(3)研究分担者
徳久雅人 (TOKUHISA MASATO)
鳥取大学・工学研究科・助教
研究者番号：10274557