様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２６　日現在

機関番号：　３２６８９
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３１７

研究課題名（和文）言語生産性：有効な類推関係クラスターの迅速な抽出・統計的機械翻訳でその評価

研究課題名（英文）Language productivity: fast extraction of productive analogical clusters and their evaluation using statistical machine translation

研究代表者

ＬＥＰＡＧＥ　ＹＶＥＳ（LEPAGE, YVES）

早稲田大学・理工学術院（情報生産システム研究科・センター）・教授

研究者番号：７０５７３６０８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：　本研究の目的は、１。単言語データから類推関係クラスターを構築し、２。そのクラスターから擬似パラレルコーパスを生成し、３。パラレルコーパスに追加することにより４。統計的機械翻訳（SMT）の精度を向上させる。
　そのため、様々なツールを実装し公開した。新しいデータ構造も導入した：類推関係グリッド。形態的に貧しい言語を始め形態豊かな言語を渡って様々な言語でデータを構築した：欧州連合の１１ヶ国語、中国語、日本語、また追加言語（アラビア語、グルジア語、ナバホ語、ロシア語、トルコ語）。データの一部分は公開した。
　行った実験で擬似パラレルコーパスの追加により日中SMTの翻訳精度を向上することを明らかにした。

研究成果の概要（英文）：　The goal of the project was 1/ to build tools to produce analogical clusters from monolingual data, 2/ to use such clusters in the production of quasi-parallel corpora, 3/ to use such quasi-parallel corpora in addition to parallel corpora 4/ to obtain improvements in translation accuracy in statistical machine translation (SMT).
　Tools were built and publicly released. In addition to what was announced in the research plan, a new data structure, analogical grid was introduced. Data were produced in morphologically poor to rich languages: 11 European languages (N-grams from word to 6-grams), Chinese, Japanese (short sentences of less than 30 characters for SMT experiments), and additional languages (word forms in Arabic, Georgian, Navajo, Russian, Turkish, etc.). Part of the data has been publicly released.
　Various experiments showed that it is possible to improve translation accuracy thanks to quasi-parallel data produced by analogy, and filtered, in SMT for Chinese-Japanese.

研究分野：　自然言語処理

キーワード：　自然言語処理　人工知能　データ構造　形態で豊かな言語　中国語・日本語

１．研究開始当初の背景
（１）機械翻訳システムを構築する際、統計的アプローチでもニューラルアプローチでも大規模パラレルコーパスが必要となるが、日中の場合、このようなコーパスは少ない。単言語コーパスを探っても対訳関係を持つ文は少ない。その問題を扱うため**言語生産性**に基づいて新しいパラレル文を生成する技術を提案した。

（２）言語生産性は、一部分は類推関係で説明できる現象である。例えば英語の未知単語 inexhaustivity を読む際、inactivity：active、insensitivity：sensitive 等の比例を利用し、exhaustive という単語を導き出すことによって inexhaustivity の意味は分かるようになる。また、言語生産性で生成した新しい単語「例えば inexhaustivity」を話す際、その単語の信頼性（分かり易さ）を推測する必要がある。

２．研究の目的
（ａ）単語・文のレベルともに inactivity：active、insensitivity：sensitive のような比例集合（**類推関係クラスター**と呼ぶ）を単言語コーパスから**迅速に**抽出する技術の開発。

（ｂ）類推関係クラスターを利用し、謂わば**言語生産性**に基づいて、新しい単語・文を生成する技術の開発。

（ｃ）新しい単語・文を生成する際、生成されたものの信頼性を測るテストの検討、又は信頼性の低いものを削除する技術の開発。そのため、**有効な類推関係クラスター**を位置付ける手法の検討。

（ｄ）**統計的機械翻訳**のため、類推関係クラスターを利用し、新しい文を生成し、信頼性の篩をかけ、残った文の中から対訳関係を位置づけ、パラレルコーパスを構築する方法を設計し実装する。その効果の評価。

上記述べた４つの目的を考慮し、本研究の課題名は「言語生産性：有効な類推関係クラスターの迅速な抽出・統計的機械翻訳でその評価」とした。

３．研究の方法
（ａ）を達成するため、二部分に分けた。初めに一部分で、2014 年に研究代表者が発表したアルゴリズムの**加速**を検討した。次にもう一部分はグラフ理論的な難しい問題があるため（従来のアルゴリズムの複雑**さ**は高い）、解決として、データをすべてカバーする新しい貪欲アルゴリズムを提案した。できた類推関係クラスター抽出プログラムは、様々な言語データに適応させ、速度を測った。本研究のさらなる成果の一つとして、類推関係グリッドという新しいデータ構造を提案し、そのデータ構造を類推関係クラスターから計算し、そのためのアルゴリズムを設計した。（ｃ）に関して、より有効（単語・文をどの程度で生成できるのか）な類推関係クラスター・グリッドを抽出するため、充塡率という変数の影響を測った。いずれの実験でツールとして扱うため、類推関係クラスター又はグリッド抽出プログラムをパッケージにした。（ｄ）の目的に関して、実験パイプラインを構築した：上記のプログラムで単言語データから類推関係クラスター構築、そのクラスターから擬似パラレルコーパス生成、パラレルコーパスに追加することにより翻訳精度測定。（生成されたコーパスは完全なパラレルコーパスではないため、擬似パラレルコーパスと呼ぶ。）擬似パラレルコーパスは、新しく生成された対訳関係をもつ文対の集合である。生成された文は対訳度が高い類推関係クラスターと単言語データの間の類推方程式の解になる。そのため、類推関係クラスターの対訳関係を測定する手法と類推関係クラスターでの類推方程式の解を求める手法を提案した（目的（ｂ））。

実験のため、機械を２台購入した。類推関係クラスターを構築する際、メモリ不足問題が発生したため、メモリも購入した。データの管理のため HDD を購入した。購入した機械を使用し数多くの実験を行った。形態素の実験は、まず、単語のレベルの実験では、言語族的に幅広い聖書の様々な翻訳版のデータを利用した。また、形態論の国際評価キャンペーン SIGMORPHON のデータも利用した。機械翻訳の実験は、本研究室でウェブから抽出した日中データと国際間評価キャンペーン WAT の日中データを利用した。

研究目的を達成するまで、時間がかかり、プログラムを実装するため、研究代表者だけでなく、研究協力者が参加した。実験は、研究代表者も行い、研究協力者にも依頼した。形態学的に豊かな言語であるインドネシア語の結果分析に研修生を雇った。

本研究で、様々な成果が得られたため、多数の国際発表を行った。また、３年度から、研究代表者は招待講演で成果発表ができた。

４．研究成果
（１）Python のモジュールとして様々な機能が含まれているツールを公開した。まず、基礎機能は高速編集距離計算と正確さが高い高速類推関係方程式の求解（目的（ｂ））。次に高速類推関係クラスターの構築（目的（ａ））。最後に類推関係グリッドの構築。類推関係グリッドは、本研究のさらなる成果であり、新しい語形を生成する際、重要な役割を果たすデータ構造である（目的（ｃ））。また、本研究の成果の一つとして、比例、類推

関係、類推関係クラスター、類推関係グリッドの理論的定義が明確になった。一般化ができ、公開したツールはそれを反映している。最初に形式類推関係のみ（基本的に文字数と編集距離）の類推関係に応用できた。結果的に使用者が自由に定義した単語や文の整数値ベクトル記述にも応用できた。一般化したことで、公開したツールは、世界中の研究者により多く利用されることを期待する。
【学会発表 1a，4a；7a】

（２）データを公開した。一つはヨーロッパ１１ヶ国語の類推関係クラスターと類推関係グリッドである。元のデータは統計的機械翻訳でよく使われてきた Europarl コーパスである。対訳関係を持つ最初の千文から N＝１から６までの N グラムを抽出し、それぞれの言語で類推関係クラスターと類推関係グリッドを構築したものである。このデータを公開したことで、世界中の研究者にそれぞれの言語の構造特徴の新しい観点が検討されることを期待する。もう一つは、国際評価キャンペーン SIGMORPHON のデータに基づく、１０ヶ国語（アラビア語、フィンランド語、グルジア語、ドイツ語、ハンガリー語、マルタ語、ナバホ語、ロシア語、スペイン語、トルコ語）で６千５百万個の形式類推関係集合である。世界初の大規模類推関係集合である。国際会議論文では、様々な形態論データを使用し、類推関係グリッド構築について発表した。例えば、言語族的に幅広い様々な聖書の翻訳版を使用し（４大陸から異なる３言語＝１２言語)、類推関係グリッドの大きさと充填率を検討した。類推関係グリッド生成時間は形態学的豊かさに反映することを明らかにした。また、４ヶ国語（英語、ロシア語、ギリシア語とインドネシア語）で、語彙の豊かさを測り、ある著者が使っている語形は他の著者の単語で説明可能か検討した。
【学会発表 1a，3a，5a，8a；5b】

（３）類推関係方程式の解を求める面で（目的（ｂ))、二つの研究を行った。一つは新しい手法をアルゴリズム的にも機械学習的にも提案した。二つ目は、与えられた語形や文を類推関係クラスターを全体的に使用し、新しい語形や文を生成するための手法を検討した。これは、日中統計的機械翻訳の実験で擬似パラレルコーパス生成の際利用した。
【学会発表 4a，7a；5b】

（４）統計的機械翻訳の評価では（目的（ｄ))、擬似パラレルコーパスに含まれている文は類推関係で生成された文であるため、その品質は問題となる。ここは、二つの課題を扱った。一つは、一般的に類推関係で生成される語形の信頼性の検討である。形態学的豊かな言語であるインドネシア語を例として選んだ。類推関係グリッドで生成可能な語形の数は、どの程度でグリッドの大きさと充填率に依存するかを検討した。結果は充填率が高ければ高いほど信頼性が上がる。また、生成された語形の実際の正しさを確かめるため、二つのインドネシア語形態素解析器で解析を行った。実験の結果では生成されたものは半分以上説明できた。最後に、統計的検定で信頼性を推定する検討もした。４ヶ国語（英語、ドイツ語、フィンランド語とインドネシア語）で行なった実験の結果、フィッシャーの正確確率検定の使用で、信頼性がある若干上がることが明らかになった。もう一つは、類推関係で生成された文の場合、文法・意味的な正しさを保証するための手法を検討した。二つの手法を組み合わせることで、非常に高い文法・意味的な正しさが保証できた。一つの手法は文字列のある参照コーパスで存在を確認する手法であり、もう一つは、機械翻訳での翻訳精度の尺度 BLEU に基づく手法である。
【学会発表 2a，6a，7a；1b，2b，3b，4b】

（５）統計的機械翻訳の評価に関し、予備実験で小規模コーパスの条件で未知単語列を翻訳するために類推関係の使用が可能であることを明らかにした。しかし、通常の条件では向上は見られなかった。本研究課題では、翻訳する文の単語列の翻訳、翻訳過程の最中ではなく、予めある程度で大きなパラレルコーパスの構築が目標であった。（３）で述べたように、パラレルコーパスではなく、擬似パラレルコーパスを生成する。公開された類推関係クラスター構築ツールに基づき、類推関係クラスターの対訳度の推測する手法を提案し、翻訳までのパイプラインを設計し、実装した。本研究室でウェブから抽出した日中データと国際間評価キャンペーン WAT の日中データで数多くの実験を行なった。本研究と過去数年間で得られた結果をまとめると、二つの条件の下で、統計的に有意で大きな向上を得ることができた（BLEU で+6 点)。一つめの条件は類推関係クラスターを構築する際、翻訳データと異なるデータを使用する。もう一つは比較的に大きな擬似パラレルコーパスを使用する。
【雑誌論文 1，2、招待講演 1，2，3】

（６）研究成果公開：2017 年 7 月、2017 年 12 月と 2018 年 5 月に研究代表者は３回も本研究の成果について招待講演を行った。その際、類推関係クラスター・グリッドの概念を説明し、ツールを案内し、統計的機会翻訳その評価について紹介した。

５．主な発表論文等
〔雑誌論文（査読付き)〕（計２件）

1. W. Yang, H. Shen, and Y. Lepage. Inflating a small parallel corpus into a large quasi-parallel corpus using monolingual data for Chinese-Japanese machine translation. J. of Inf. Proc., 25:88-99, 2017. DOI:10.2197/ipsjjip.25.88

2. J. Luo and Y. Lepage. A method of generating translations of unseen n-grams by using proportional analogy. IEEJ Transactions in Electronics, Information and Systems, 11(3):325-330, May 2016. DOI:10.1002/tee.22221［本論文は本研究課題の予備実験の結果である］

〔学会発表（査読付き）〕（計８件）

1a. R. Fam and Y. Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In LREC 2018, p. 1060-1066, Miyazaki, May 2018.
URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/344.pdf

2a. R. Fam, A. Purwarianti, and Y. Lepage. Plausibility of word forms generated from analogical grids in Indonesian. In ICCA 2018, p. 179-184, Yangon, Feb. 2018.

3a. R. Fam and Y. Lepage. A holistic approach at a morphological inflection task. In LTC'17, p. 88-92, Poznan, Nov. 2017.
URL: http://ltc.amu.edu.pl/book/papers/MOR-1.pdf

4a. Y. Lepage. Character-position arithmetic for analogy questions between word forms. In ICCBR-17, p. 17-26, Trondheim, August 2017.
URL: https://pdfs.semanticscholar.org/2214/9f2150253f5a2d02ade5dba801a4a99cedfd.pdf

5a. R. Fam and Y. Lepage. A study of the saturation of analogical grids agnostically extracted from texts. In ICCBR-17, p. 7-16, Trondheim, Aug. 2017.
URL: http://ceur-ws.org/Vol-2028/paper1.pdf

6a. R. Fam, Y. Lepage, S. Gojali, and A. Purwarianti. A study in explaining unseen words in Indonesian using analogical clusters. In ICCA 2017, p. 416-421, Yangon, Jan. 2017.

7a. V. Kaveeta and Y. Lepage. Solving analogical equations between strings of symbols using neural networks. In ICCBR-16, p. 67-76, Atlanta, Oct. 2016.
URL: http://ceur-ws.org/Vol-1815/paper7.pdf

8a. R. Fam and Y. Lepage. Morphological predictability of unseen words using computational analogy. In ICCBR-16, p. 51-60, Atlanta, Oct. 2016.
URL: http://ceur-ws.org/Vol-1815/paper5.pdf

〔学会発表（査読なし）〕（計５件）

1b. R. Fam and Y. Lepage. Validating analogically generated Indonesian words using Fisher's exact test. In 言語処理学会第２４年次大会論文集, p. 312-315, Okayama, March 2018.
URL: http://anlp.jp/proceedings/annual_meeting/2018/pdf_dir/C2-1.pdf

2b. F. Rashel, A. Purwarianti, and Y. Lepage. Plausibility of word forms generated from analogical grids on Indonesian. In ISIPS 2017, p. 245-247, Kitakyushu, Nov. 2017.

3b. P. Liu and Y. Lepage. Confidence of word forms generated in analogical grids. In ISIPS 2017, p. 238-240, Kitakyushu, Nov. 2017.

4b. R. Fam, Y. Lepage, S. Gojali, and A. Purwarianti. Indonesian unseen words explained by form, morphology and distributional semantics at the same time. In 言語処理学会第２３年次大会論文集, p. 178-181, Tsukuba, March 2017.
URL: http://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/P4-1.pdf

5b. W. Yang, M. Gao, and Y. Lepage. Production of analogical clusters between marker-based chunks in Chinese and Japanese. In ISIPS 2016, p. 238-241, Kitakyushu, Nov. 2016.

〔招待講演〕（計３件）

1. Y. Lepage. Quasi-parallel corpora: Hallucinating translations for the Chinese-Japanese language pair. BUCC workshop colocated with LREC 2018, no page number, Miyazaki, May 2018.
URL: http://lrec-conf.org/workshops/lrec2018/W8/pdf/11_W8.pdf

2. Y. Lepage. Automatic production of quasi-parallel corpora for machine translation. In ICNLSSP 2017, Casablanca, 06--07 Dec. 2017.
Video: https://www.youtube.com/watch?v=rELErObCw5M&feature=youtu.be
Slides: https://drive.google.com/file/d/1pbo9H7hk7TDOwiyqt91qI-6vVmgarMar/view

3. Y. Lepage. Clusters et grilles analogiques：validation par la traduction automatique. 40 ans de TA, Grenoble, France, July 2017.
Slides: http://40ansdeta.imag.fr/wp-content/uploads/2017/07/YvesLepage_40ansdeTA.pdf

〔その他〕
ホームページ：
http://lepage-lab.ips.waseda.ac.jp/
> Projects > Kakenhi 15K00317

６．研究組織

(1)研究代表者
ルパージュ・イヴ（LEPAGE, Yves）
早稲田大学・大学院情報生産システム研究科・教授
研究者番号: 70573608

(4)研究協力者
楊 巍（YANG, Wei）
ファム　ラシェル（FAM, Rashel）
スサンティ・ゴジャリ（SUSANTI GOJALI）