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１６日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：19500137
研究課題名（和文）　多言語対訳コーパスを用いた言語間距離の計算とその応用

研究課題名（英文）　Language distance and its application by using multi-lingual parallel corpus

研究代表者
隅田　英一郎（SUMITA EIICHIRO）
独立行政法人情報通信研究機構・知識創成コミュニケーション研究センター 言語翻訳グループ・グループリーダー
研究者番号：90395020

研究成果の概要（和文）：構文、換言の利用、多言語向き形態素解析等、翻訳の高度化を行い、翻訳品質評価に基づく言語間距離を計算する方式を提案した。「英語話者の学習時間」は、フランス語などは短く、アラビア語、中国語、日本語は長いことは提案距離で説明できる。しかし、後者の 3 言語の「学習時間」は同じであり、英語との距離差では説明できない。より精緻な距離の創出が今後の課題である。また、副産物として 21 言語の全組合せ 420 通りの翻訳システムを構築した。

研究成果の概要（英文）：Based on improvement in translation method using syntax, paraphrase, segmentation oriented to multi-language, we propose the method calculating a distance between languages. Although the proposed distance roughly correlates time for English speaker to acquire the concerned language, further refinement of distance definition should be pursued. As byproduct, we built translation systems for 420 directions, i.e., all possible directions among 21 languages.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,700,000	510,000	2,210,000
２００８年度	1,500,000	450,000	1,950,000
２００９年度	400,000	120,000	520,000
年度			
年度			
総　計	3,600,000	1,080,000	4,680,000

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：自然言語処理

１．研究開始当初の背景

類型論で同じ語族と分類される場合でも類似していない言語対もあれば、違う語族でも類似している場合もある。類型論では定性的な比較は可能だが、定量的な比較が出来ないなどの課題があり、これを補完する計算可能

な手法の創出が必要であった。

２．研究の目的
本研究では、対訳コーパスに基づいて言語間の相違をモデル化し、言語間の距離を計算する手法を提案する。

３．研究の方法
対訳コーパス（文単位で対応がついた対訳データ）を利用することによって計算可能となる 2 言語間の距離を提案し、これをソフトウエアとして実現する。

４．研究成果
（1）19 年度
BTEC コーパスを使って統計的翻訳システムを構築し、その翻訳性能に基づく言語間距離を検討した。この距離において、日本語と韓国語の距離が、ポルトガル語とその方言であるブラジル・ポルトガル語との距離より小さいという興味深い実験結果が得られた。
そのほか、対訳コーパス構築の手法、自動評価手法、HINDI 語など語形変化が多い言語のコーパスにおける未知語処理、翻訳モデルのクラスタリングなど翻訳方式における成果も多数上げることができた。

（2）20 年度
言語間距離の計算方法を拡張するために、以下の特徴を用いることを提案した: PIVOT 言語（翻訳時に原言語から目的言語へ直接翻訳するのでなく、第 3 の言語を仲介とする方式があり、仲介言語を PIVOT 言語と呼ぶ）を用いた時の翻訳品質、PIVOT 方式を変えた時の翻訳品質、構文利用した翻訳方式での翻訳品質、翻字の性能、フレーズテーブルの統計量、語順変更の統計量、語彙量、形態素の多様性、文法（一致の有無、SVO パターンなど）。
そのほか、コーパス構築の手法、自動評価手法、翻字による未知語処理、PIVOT 翻訳構文の利用など翻訳方式における成果も多数上げることができた。構文の利用について説明する。統計的翻訳において、語順の誤りが深刻な課題となっており、Inversion transduction grammar (ITG)などの制約が既に提案されている。ITG では, 目的言語側の語順は原言語側の可能なバイナリ木を回転に限定する。構文解析を利用して、このバイナリ木の曖昧性を解消すれば、さらに強い制約となり、探索誤りを削減できる。例えば、4 語の原言語文{ａｂｃｄ}に対して, ITG では 22 通りの語順を考慮する必要があるが、バイナリ木が ((a b) (c d))とすると、8 通りだけで済む。英中の実験では、文字の BLEU-4 で 35.2 から 37.0 に 1.8 ポイントの改善、文字誤り率で 74.1 から 67.9 に 6.2 ポイント削減できた。

（3）21 年度
構文利用の翻訳方式の提案・実装、双方向翻訳方式の提案と実装、対訳に基づいて学習する形態素解析手法の提案と実装、換言による翻訳方式の提案と実装、動的モデル統合法の提案と実装、双方向翻字の提案と実装など、翻訳の基本方式の改良を行い、多数の言語対によって、翻訳手法の汎用性を確認した。これに基づいて言語間距離を計算する基本方式を拡張した。
学習時間と現在まで考案した言語間距離（翻訳困難度）の関係について次のことが分かった。
米国の調査「英語話者が外国語の集中コースによって上級レベルに到達するまでの学習時間」は、（１）フランス語、ドイツ語、スペイン語は 20 週、（２）アラビア語、中国語、日本語は 44 週となっている。英語と（１）の言語の言語間距離（翻訳困難度）は、英語と（２）の言語の言語間距離（翻訳困難度）より小さいので学習時間とおおむね相関する。ただ、アラビア語、中国語、日本語は同じ（２）の学習時間であるが、言語間距離に関して、アラビア語、中国語、日本語の順に距離が大きく異なったので、学習時間との関係を説明するためには、さらに、言語間距離の計算手法の改善を要することがわかった。
副産物として、21 言語（英、 独、 デンマーク、 オランダ、 仏、 イタリア、 スペイン、 ポルトガル、 ブラジル、日本、 北京、韓国、 ロシア、 アラビア、 ヒンディ、インドネシア、 マレー、 タイ、 タガログ、ベトナム、 台湾華語）の翻訳システムを構築し、ネットワーク経由で iPhone で利用するシステムを構築した。

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６件）

① A. Finch、隅田英一郎、Class-dependent Modeling for Dialog Translation、IEICE TRANSACTION on Information Systems、IEICE TRANSACTIONS on Information and Systems、査読有、Vol. E92-D、No.12、2009、2378－2385
② 橋本佳、山本博史、大熊英男、隅田英一郎、徳田恵一、A Reordering Model Using a Source-Side Parse-Tree for

Statistical Machine Translation、IEICE TRANSACTIONS on Information and Systems、査読有、Vol. E92-D、No.12、2009、2386～2393

③ 山本博史、大熊英男、隅田英一郎、Imposing Constraints from the Source Tree on ITG Constraints for SMT、IEICE TRANSACTIONS on Information and Systems、査読有、Vol. E92-D、No.9、2009、1762－1770

④ 山本博史、隅田英一郎、Bilingual Cluster Based Models for Statistical Machine Translation 、IEICE TRANSACTIONS on Information and Systems、査読有、Vol. E91-D、No.3、2008、588－597

⑤ 安田圭志、隅田英一郎、機械翻訳の研究・開発における翻訳自動評価技術とその応用、人工知能学会誌、査読有、23 巻、1号、2008、2-9

⑥ 竹澤寿幸、菊井玄一郎、水島昌英、隅田英一郎、Multilingual Spoken Language Corpus Development for Communication Research、査読有、Vol. 12、No. 3、2007、303-324

〔学会発表〕(計１６件)

① A. Finch、隅田英一郎、Transliteration by Bidirectional Statistical Machine Translation、The 2009 Named Entities Workshop, ACL-IJCNLP 2009、2009 年 8 月 7 日、Suntec (シンガポール)

② A. Finch、隅田英一郎、Bidirectional Phrase-based Statistical Machine Translation、The 2009 Conference on Empirical Methods in Natural Language Processing、2009 年 8 月 6 日、Suntec (シンガポール)

③ 橋本佳、山本博史、大熊英男、隅田英一郎、徳田恵一、Reordering Model Using Syntactic Information of a Source Tree for Statistical Machine Translation、SSST-3, Third Workshop on Syntax and Structure in Statistical Translation、2009 年 6 月 5 日、University of Colorado at Boulder (米国)

④ M. Paul、山本博史、大熊英男、隅田英一郎、中村哲、On the Importance of Pivot Language Selection for Statistical Machine Translation、NAACL HLT 2009、2009 年 6 月 2 日、University of Colorado at Boulder (米国)

⑤ M. Paul、A. Finch、隅田英一郎、NICT@WMT09: Model Adaptation and Transliteration for Spanish-English SMT 、EACL 2009 Fourth Workshop on Statistical Machine Translation、2009 年 3 月 30 日、Athens (ギリシャ)

⑥ 山本博史、大熊英男、隅田英一郎、Imposing Constraints from the Source Tree on ITG Constraints for SMT 、ACL-08:HLT SSST-2 (The Second Workshop on Syntax Structure in Statistical Translation)、2008 年 6 月 20 日、Columbus (米国)

⑦ K. Arora 、M. Paul 、隅田英一郎、Translation of unknown words in phrase-based Statistical Machine Translation for languages of rich morphology 、SLTU-2008 (The first International Workshop on Spoken Languages Technologies for Under-resourced Languages)、2008 年 5 月 6 日、Hanoi University of Technology (ベトナム)

〔産業財産権〕
○出願状況 (計２件)

名称：機械翻訳装置、機械翻訳方法、およびプログラム
発明者：大西貴士、内山将夫、隅田英一郎
権利者：(独)情報通信研究機構
種類：特許
番号：特願 2010-044213
出願年月日：２２年３月１日
国内外の別：国内

名 称 : Language Independent Word Segmentation for Statistical Machine Translation
発明者：M. Paul、A. Finch、隅田英一郎
権利者：(独)情報通信研究機構
種類：特許
番号：特願 2009-273137
出願年月日：２１年１２月１日
国内外の別：国内

○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：

６．研究組織
(1)研究代表者

隅田　英一郎（SUMITA　EIICHIRO）
独立行政法人情報通信研究機構・知識創成コミュニケーション研究センター 言語翻訳グループ・グループリーダー
研究者番号：90395020

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
山本　博史（YAMAMOTO　HIROFUMI）
近畿大学・理工学部・教授
研究者番号：00395013
（H19：研究分担者）
パウル　ミヒャエル（PAUL　MICHAEL）
独立行政法人情報通信研究機構・知識創成コミュニケーション研究センター言語翻訳グループ・専攻研究員
研究者番号：20395031