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

令和　４ 年　 ９ 月　 ５ 日現在

機関番号：３４３１５
研究種目：若手研究(A)
研究期間：2017～2020
課題番号：１７Ｈ０４７０６

研究課題名（和文）低資源言語のための言語資源作成サービスネットワークの構築

研究課題名（英文）Language Services Network for Bilingual Dictionary Creation in Low Resource Languages

研究代表者

村上　陽平（Murakami, Yohei）

立命館大学・情報理工学部・准教授

研究者番号：００４３５７８６

交付決定額（研究期間全体）：（直接経費）　　18,900,000円

研究成果の概要（和文）：本研究は、言語資源の少ない低資源言語を対象に、近縁言語間の対訳辞書を網羅的に生成するための言語資源生成サービスネットワークを構築することを目的とした。基盤研究では、ピボット言語を介して二つの対訳辞書から対象言語間の対訳辞書を高精度に生成する対訳辞書の帰納的生成手法を考案した。また、複数の対象言語の対訳辞書を効率的に生成するために、帰納的生成手法と人手による作成を組み合わせた最適プランを算出するアルゴリズムを考案した。実証研究では、基盤研究の成果を実装したクラウドソーシングのための協調作業支援サービスを構築し、7つのインドネシアの民族語の21編の対訳辞書を作成した。

研究成果の学術的意義や社会的意義
辞書やコーパス、機械翻訳などの言語資源の偏在が、言語資源の少ない低資源言語の問題を生じさせ、母語の異なる話者間でデジタルデバイドを引き起こしている。特に、多様な言語の存在するアジアにおいて顕著である。本研究は、このような社会的な問題に対して、低資源言語の対訳辞書を網羅的に生成する手法を提案し、インドネシアの民族語での実証を通して大きく貢献している。さらに、本研究は、言語資源作成時の品質だけでなく、費用対効果という新しい観点も導入することで、人手の作成作業を避けることのできない低資源言語の言語資源生成を大規模化へと発展させるものである。

研究成果の概要（英文）：The purpose of this study is to develop a language resource creation service network to comprehensively create bilingual dictionaries between closely-related languages for low-resource languages. In the basic research, we proposed a bilingual dictionary induction that generated a bilingual dictionary between target languages by combining two bilingual dictionaries via a pivot language. We also proposed an algorithm to calculate the optimal plan consisting of the dictionary induction and manual creation with the lowest cost. In the empirical research, we developed a collaborative dictionary creation support service for crowdsourcing by implementing the outcomes of the basic research and generated 21 bilingual dictionaries among 7 Indonesian ethnic languages.

研究分野：サービスコンピューティング

キーワード：サービスコンピューティング　Webサービス　低資源言語　言語資源　クラウドソーシング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

辞書やコーパス、機械翻訳などの言語資源の偏在が、言語資源の少ない低資源言語の問題を生じさせ、母語の異なる話者間でデジタルデバイドを引き起こしている。欧州言語資源協会(ELRA)が管理している、言語資源のデータベース LRE Map の統計データによると、100 以上の資源が存在する言語が 11 言語しかなく、その話者人口は地球全体の人口の 47%程度である。

低資源言語の言語資源獲得は、言語資源分野において重要な課題である。ACM はアジア言語を対象としていた論文誌を低資源言語にまで拡大して、Transactions on Asian and Low-Resource Language Information Processing (TALLIP)に名称を変更しており、分野としても立ち上がり始めている。この分野の特徴は、大量のテキストデータを統計的に処理することでモデルを構築している現在の自然言語処理とは異なり、データが限られて小規模な点である。そのため，既存の言語資源や人手によるデータ構築と組み合わせることでデータの不足を補いながら、モデルを構築する手法が求められている。

２．研究の目的

本研究の目的は、サービスコンピューティングの技術を用いて、帰納的な辞書生成プログラムと人手による辞書作成や評価タスクを連携させて、低資源言語の言語資源を生成するサービスネットワークを構築することである。

そのためには、まず、(1)既存の限られた対訳辞書を組み合わせて新規の対訳辞書を帰納的に生成するメカニズムを考案する必要がある。次に、同一言語族の近縁言語間の対訳辞書を網羅的に生成するために、(2)対訳辞書の生成プランニングを行い、生成コストを最小化する対訳辞書の生成順序を同定する必要がある。

さらに、生成された対訳辞書の精度を継続して高めるためのオンラインメカニズムを構築し、実証の場に適用することで、低資源言語の対訳辞書を継続して作成し、研究成果の評価を行う。

３．研究の方法

本研究は、研究室における基盤研究と現場における実証研究を、当初より並行して実施する。特に、人手による対訳辞書の作成は、少数の対象言語から始め、継続して言語数や辞書サイズを拡大していくことで、実証研究で得られたデータや知見を基盤研究の問題設定にその都度反映し、フィードバックループを形成しながら進める。そのため、研究開始当初から、現場となるインドネシアの大学と連携し、実証研究を実施する現場の確保と拡大に努める。なお、生成された低資源言語の対訳辞書を言語グリッドを通じて世界に公開し、言語資源の蓄積も同時に図っていく。

本研究では、人間とプログラムの協働による低資源言語の言語資源生成を実現するために、以下の 3 課題に取り組む。

(1) 対訳辞書の帰納的生成
二つの対訳辞書を共通の言語（ピボット言語）で連結し、生成されたグラフから一つの対訳辞書を帰納的に生成する。具体的には、同族言語間の語義の類似性に基づいて、意味的制約を導入し、対訳辞書の生成をこれらの制約に基づく最適化問題として定式化して単語の対訳関係を抽出する。対象とする言語間の類似度に応じて意味的制約を変更し最適化問題を解くことで、従来の逆引き手法が達成していた適合率を維持しつつ、再現率を向上させる。

(2) 対訳辞書の生成プランニング
同言語族の近縁言語全体の対訳辞書生成コストを最小化するには、どの言語間の対訳辞書からどの手法で作成していくのかプランニングすることが極めて重要である。自動生成技術では言語間の類似度や入力辞書のサイズによって生成される対訳辞書の精度やサイズが異なってくるため、このプランニングは状態遷移に不確実性を伴う意思決定の連続となる。そこで、マルコフ決定過程に基づく対訳辞書生成プランニングに取り組み、環境に動的に適応しながら対訳辞書サービスを連携し、作成コストを最小化する。

(3) クラウドソーシングによる対訳辞書生成
対訳辞書の帰納的生成手法と生成プランニングをクラウドソーシング用の Web サービスに組み込み、対訳辞書生成の協働作業に適用し、インドネシアの民族言語 10 言語程度の対訳辞書

を作成する。各言語を理解する作業者が必要なため、インドネシアの現地の大学と連携し、作業者のネットワークを拡大する。

４．研究成果

基盤研究では、(1)対訳辞書の帰納的生成、(2)対訳辞書の生成プランニングの 2 課題に取り組んだ。実証研究では、(3)クラウドソーシングによる対訳辞書生成支援システムを実装した。以下、それぞれの課題について、得られた成果を述べる。

(1) 対訳辞書の帰納的生成

同根語のみを対訳として抽出する、既存の対訳辞書の帰納的生成手法は再現率が低い。そこで、再現率を向上させるために、従来研究によって定式化された制約最適化アルゴリズムを一般化し制約を緩めることで、同根語以外の対訳を獲得する手法を考案した。まず、抽出された対訳ペアを基に繰り返し制約最適化を計算するようフレームワークを拡張した。さらに、ピボット単語の多義性や、正書法に基づく単語の綴りの類似度など複数のヒューリスティックスをコストの重み付け関数として定義し、同根語以外の対訳を獲得した。この結果、インドネシア語、マレー語、ミナンカバウ語やドイツ語、英語、オランダ語など４種類の同言語族の３言語組に対して、従来手法よりも平均で適合率を 34%程度下げつつも、再現率は 183%程度向上させており、提案手法の有効性を示した。特に、3 言語間の言語間類似度が平均 73%と高いインドネシア語、マレー語、ミナンカバウ語においては、適合率の低下を 11.5%程度に抑えつつ、再現率を 191%向上し、F 値は 72%向上させている。

このように、対訳辞書の帰納的生成は対象言語間が類似しているほど効果が高いため、効果的に多くの対訳辞書を生成するために、言語間類似度データベースから近縁言語のクラスタを抽出する手法を考案した。具体的には、階層的クラスタリングにより、類似度が閾値を超える言語のみからなる緊密なクラスタを抽出し、この緊密クラスタを基準に k-平均法により全体を分割していくことで、できる限り安定度の高いクラスタの集合を獲得する。提案手法を 32 のインドネシアの民族言語に実際に適用し、マレー系の言語クラスタやバタック系の言語クラスタなど５つの近縁言語クラスタを同定できることを確認した。

これらの成果を、論文誌の ACM Transactions on Asian and Low-Resource Language Information Processing や International Journal of Electrical and Computer Engineering などに発表した。

(2) 対訳辞書の生成プランニング

低資源言語では、ピボット言語に基づく対訳辞書の帰納的生成手法の入力となる対訳辞書が十分に存在せず、人手による対訳辞書の作成が必要となることがある。総作業コストを低減するには、人手による対訳辞書の作成と、ピボット言語に基づく対訳辞書の帰納的生成をどのように組み合わせるかというプランニングの問題となる。そこで、この問題をマルコフ決定問題として定式化し、対訳辞書の生成プランを算出するアルゴリズムを考案した。このアルゴリズムによって導出された最適プランを評価するために、インドネシア大学、テレコム大学、イスラミックリアウ大学の協力のもと、インドネシア語、マレー語、スンダ語、ジャワ語、ミナンカバウ語の５つのインドネシアの民族言語を対象に評価実験を行った結果、5 言語分の 10 編の対訳辞書を作成するのに、全て人手で作成するよりも、42%コストを削減できることを確認した。さらに、実際に要した所要時間と、対訳辞書生成プランの推定所要時間の差は 3%程度であり、提案手法が高精度で総コストを推定できることを確認した。

これらの成果を、論文誌では ACM Transactions on Asian and Low-Resource Language Information Processing や Journal of Data Science and Its Applications、国際会議では LREC などに発表した。

(3) クラウドソーシングによる対訳辞書生成

対訳辞書生成プランニングで算出したプランに従って、人手による辞書作成や評価を行えるように、クラウドソーシングのための協調作業支援サービスを構築した。これにより、対訳辞書の対象言語のバイリンガルでなくても、ピボット言語を介して対訳ペアの作成や評価を行うことを可能にしている。

また、実証段階では、一度に多くの言語の作業者を確保できず、段階的に対象言語を拡張していかなければならないことがあるため、初期状態から目標状態までの最適プランを一度に計算できない。そこで、対訳辞書の生成プランニングを対訳辞書の作成状態に基づく動的な再プランニ

ングに拡張した。具体的には、正しいと評価された対訳ペア数に基づいてプランニングの探索空間を随時絞り込むことで、ポリシーの再計算を行いプランの最適化を行う。実際に提案手法を用いて、インドネシア語、マレー語、ミナンカバウ語、ジャワ語、スンダ語の辞書を作成した後に、バンジャル語とパレンバン語を追加して合計 21 編の対訳辞書を作成し、人手で作成するよりも約 60%のコストを削減した。

さらに、人手による対訳の評価タスクの精度を向上させるために、協調作業支援サービスに超問題を用いた評価結果の集約手法を適用した。実際に、ミナンカバウ語の能力を 5 段階に分けたインドネシア人被験者 20 人から得られた作業結果に基づき、作業者の回帰モデルを構築し評価を行った。提案手法をインドネシア語とミナンカバウ語の 1000 語の対訳辞書作成に適用することで、提案手法はいずれの作業者の能力平均においても、多数決よりも対訳辞書の正確性を向上させることを確認した。

これらの成果を、雑誌では IEEE Computer、論文誌では Information、ヒューマンインタフェース学会論文誌、国際会議では CollabTech、LREC などに発表した。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　8件／うち国際共著　4件／うちオープンアクセス　5件）

１．著者名 Arbi Haza Nasution, Yohei Murakami, Toru Ishida	４．巻 vol. 20, No. 2
２．論文標題 Plan Optimization to Bilingual Dictionary Induction for Low-resource Language Families	５．発行年 2021年
３．雑誌名 ACM Transactions on Asian and Low-Resource Language Information Processing	６．最初と最後の頁 29:1-29:28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3448215	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 西村一球, 村上陽平, Pituxcoosuvarn Mondheera	４．巻 Vol. 23, No. 2
２．論文標題 画像特徴量に基づく文化差検出	５．発行年 2021年
３．雑誌名 ヒューマンインタフェース学会論文誌	６．最初と最後の頁 145-152
掲載論文のDOI（デジタルオブジェクト識別子） 10.11184/his.23.2_145	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Donghui Lin, Yohei Murakami, Toru Ishida	４．巻 Vol. 11, No. 2
２．論文標題 Towards Language Service Creation and Customization for Low-Resource Languages	５．発行年 2020年
３．雑誌名 Information	６．最初と最後の頁 67
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/info11020067	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Arbi Haza Nasution, Yohei Murakami	４．巻 Vol. 2, No. 2
２．論文標題 Visualizing Language Lexical Similarity Clusters: A Case Study of Indonesian Ethnic Languages	５．発行年 2019年
３．雑誌名 Journal of Data Science and Its Applications (JDSA)	６．最初と最後の頁 50-60
掲載論文のDOI（デジタルオブジェクト識別子） 10.34818/jdsa.2019.2.23	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Yohei Murakami, Takao Nakaguchi, Donghui Lin, Toru Ishida	4．巻 11422
2．論文標題 Two-Layer Architecture for Distributed Massively Multi-agent Systems	5．発行年 2019年
3．雑誌名 Massively Multi-Agent Systems II	6．最初と最後の頁 53-65
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-20937-7_4	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yohei Murakami	4．巻 1192
2．論文標題 Indonesia Language Sphere: an ecosystem for dictionary development for low-resource languages	5．発行年 2019年
3．雑誌名 Journal of Physics: Conf. Series	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1088/1742-6596/1192/1/012001	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Arbi Haza Nasution, Evizal Abdul Kadir, Yohei Murakami, Toru Ishida	4．巻 -
2．論文標題 Toward Formalization of Comprehensive Bilingual Dictionaries Creation Planning as Constraint Optimization Problem	5．発行年 2020年
3．雑誌名 Optimization Based Model Using Fuzzy and Other Statistical Techniques Towards Environmental Sustainability	6．最初と最後の頁 41-54
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-15-2655-8_3	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Arbi Haza Nasution, Yohei Murakami, Toru Ishida	4．巻 Vol. 9, No. 1
2．論文標題 Generating similarity cluster of Indonesian languages with semi-supervised clustering	5．発行年 2019年
3．雑誌名 International Journal of Electrical and Computer Engineering	6．最初と最後の頁 pp. 531-538
掲載論文のDOI（デジタルオブジェクト識別子） 10.11591/ijece.v9i1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Toru Ishida, Yohei Murakami, Donghui Lin, Takao Nakaguchi, Masayuki Otani	4．巻 Vol. 51, Issue 6
2．論文標題 Language Service Infrastructure on the Web: The Language Grid	5．発行年 2018年
3．雑誌名 IEEE Computer	6．最初と最後の頁 pp. 72-81
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/MC.2018.2701643	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kemas M. Lhaksmana, Yohei Murakami, Toru Ishida	4．巻 Vol. 28, No. 1
2．論文標題 Role-Based Modeling for Designing Agent Behavior in Self-Organizing Multi-Agent Systems	5．発行年 2018年
3．雑誌名 International Journal of Software Engineering and Knowledge Engineering	6．最初と最後の頁 79-96
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0218194018500043	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Arbi Haza Nasution, Yohei Murakami, Toru Ishida	4．巻 Vol. 17, No. 2
2．論文標題 A Generalized Constraint Approach to Bilingual Dictionary Induction for Low-Resource Language Families	5．発行年 2018年
3．雑誌名 ACM Transactions on Asian and Low-Resource Language Information Processing	6．最初と最後の頁 9:1-9:28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3138815	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計17件（うち招待講演　3件／うち国際学会　11件）

1．発表者名 Ikkyu Nishimura, Yohei Murakami, Mondheera Pituxcoosuvarn
2．発表標題 Image-Based Detection Criteria for Cultural Differences in Translation
3．学会等名 26th International Conference on Collaboration Technologies and Social Computing (CollabTech 2020)（国際学会）
4．発表年 2020年

１．発表者名
Mondheera Pituxcoosuvarn, Yohei Murakami, Donghui Lin, Toru Ishida

２．発表標題
Effect of Cultural Misunderstanding Warning in MT-Mediated Communication

３．学会等名
26th International Conference on Collaboration Technologies and Social Computing (CollabTech 2020)（国際学会）

４．発表年
2020年

１．発表者名
元澤海月，村上陽平，Mondheera Pituxcoosuvarn

２．発表標題
児童の異文化コラボレーションにおけるファシリテーションの分析

３．学会等名
電子情報通信学会ヒューマンコミュニケーション基礎研究会

４．発表年
2020年

１．発表者名
張 禹王，村上陽平

２．発表標題
低資源言語ためのブロックチェーンに基づく非中央集権型辞書システム

３．学会等名
電子情報通信学会サービスコンピューティング研究会

４．発表年
2021年

１．発表者名
松岡勇樹，村上陽平

２．発表標題
コグニティブサービスを用いた翻訳エージェント

３．学会等名
電子情報通信学会サービスコンピューティング研究会

４．発表年
2021年

１．発表者名
地田大樹，村上陽平

２．発表標題
対訳辞書作成のための信頼に基づくクラウドソーシングの評価

３．学会等名
電子情報通信学会サービスコンピューティング研究会

４．発表年
2020年

１．発表者名
大久保弘基，村上陽平

２．発表標題
グラフ埋め込みを用いた代替サービスの推薦

３．学会等名
電子情報通信学会サービスコンピューティング研究会

４．発表年
2020年

１．発表者名
Yohei Murakami

２．発表標題
Language Sphere: A Socio-Technical Approach to Bilingual Dictionary Creation for Indigenous Languages

３．学会等名
International Conference Language Technologies for All（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Yohei Murakami

２．発表標題
Indonesia Language Sphere: an ecosystem for dictionary development in low-resource languages

３．学会等名
The 2nd International Conference on Data and Information Science (ICoDIS 2018)（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Arbi Haza Nasution, Yohei Murakami

２．発表標題
Visualizing language lexical similarity clusters: a case study of Indonesian ethnic languages

３．学会等名
The 2nd International Conference on Data and Information Science (ICoDIS 2018)（国際学会）

４．発表年
2018年

１．発表者名
Arbi Haza Nasution, Yohei Murakami, Toru Ishida

２．発表標題
Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages

３．学会等名
11th edition of the Language Resources and Evaluation Conference (LREC 2018)（国際学会）

４．発表年
2018年

１．発表者名
山内翔大，村上陽平，中口孝雄，石田亨

２．発表標題
ユーザ辞書を用いたニューラル翻訳のカスタマイゼーション

３．学会等名
2018年度 人工知能学会全国大会（第32回）

４．発表年
2018年

１．発表者名
Yohei Murakami

２．発表標題
The Language Grid:Towards a Worldwide Language Service Infrastructure

３．学会等名
The International Conference on Science Enginering and Technology (ICoSET 2017)（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
Arbi Haza Nasution, Yohei Murakami, Toru Ishida

２．発表標題
Similarity Cluster of Indonesian Ethnic Languages

３．学会等名
The International Conference on Science Enginering and Technology (ICoSET 2017) (国際学会)

４．発表年
2017年

１．発表者名
Takao Nakaguchi, Yohei Murakami, Donghui Lin, Toru Ishida

２．発表標題
Federation of Language Service Infrastructures for Global Collaboration

３．学会等名
The International Conference on Culture and Computing (Culture and Computing 2017) (国際学会)

４．発表年
2017年

１．発表者名
Arbi Haza Nasution, Yohei Murakami, Toru Ishida

２．発表標題
Plan Optimization for Creating Bilingual Dictionaries of Low-Resource Languages

３．学会等名
The International Conference on Culture and Computing (Culture and Computing 2017) (国際学会)

４．発表年
2017年

１．発表者名
Junta Koyama, Yohei Murakami, Donghui Lin

２．発表標題
Situated Sensor Composition For Event-Based System

３．学会等名
The 14th IEEE International Conference on Services Computing (IEEE SCC 2017) (国際学会)

４．発表年
2017年

〔図書〕　計1件

1．著者名	4．発行年
Yohei Murakami, Donghui Lin and Toru Ishida	2018年
2．出版社	5．総ページ数
Springer Singapore	225
3．書名	
Services Computing for Language Resources	

〔産業財産権〕

〔その他〕

インドネシア言語スフィア
http://langsphere.org

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
インドネシア	Islamic University of Riau	Telkom University	University of Indonesia	