Bilingual Lexicon Induction Framwork for Closely Related Languages(Abstract_要旨)

AUTHOR(S):

Arbi, Haza Nasution

CITATION:

Arbi, Haza Nasution. Bilingual Lexicon Induction Framwork for Closely Related Languages. 京都大学, 2018, 博士(情報学)

ISSUE DATE:

2018-09-25

URL:

https://doi.org/10.14989/doctor.k21395

（続紙 1 ）

京都大学	博士（情報学）	氏名	Arbi Haza Nasution (アルビ ハザ ナステイオン)
論文題目	Bilingual Lexicon Induction Framework for Closely Related Languages （近縁言語のための帰納的な対訳辞書生成フレームワーク）		

（論文内容の要旨）

The objective of this thesis is to support the development of comprehensive sets of bilingual dictionaries among closely related low-resource languages. To this end, it generalizes the one-to-one bilingual lexicon induction method to obtain many-to-many translation pairs. Furthermore, to reduce the total cost of bilingual dictionary creation, it combines machine and manual creations and optimizes creation orders. This thesis is organized into seven chapters.

Chapter 1 outlines this thesis, including research objectives, issues, and a summary of the proposed solutions.

Chapter 2 presents the background of this thesis by reviewing previous bilingual lexicon induction methods for low-resource languages. Three major methods are reviewed: extraction of bilingual lexicons from comparable corpora, pivot-based induction of new bilingual lexicons from existing ones, and an orthographic method.

Chapter 3 proposes a method for generating language similarity clusters so as to select closely related languages. To this end, the ASJP (Automated Similarity Judgment Program) database is utilized to obtain language similarity between all pairs of languages. Hierarchical clustering is then applied to identify dense clusters, where the similarity between any two languages in a cluster is higher than a predefined threshold. This method is applied to Indonesian ethnic languages to select closely related languages for the following chapters.

Chapter 4 presents a constraint-based method to induce translation pairs. Bilingual dictionaries between closely-related languages (target languages, hereafter) are induced by combining two bilingual dictionaries, each of which consists of one of the target languages and a common pivot language. Since the existing constraint-based method only extracts cognates (words with a common etymological origin) as translation pairs, its recall rate is low. To increase the recall rate, the method is generalized to obtain cognate synonyms by relaxing constraints in various ways. The result is that the generalized method (64% average F-score) significantly outperforms the existing method (41% average F-score).

Chapter 5 proposes an algorithm to optimize the process of creating a set of bilingual dictionaries among closely related languages. If the languages are low-resourced, there may not exist enough source bilingual dictionaries to perform machine creation. Manual creations of translation pairs are performed in that case. To optimally combine machine and manual

creations, dictionary creation planning is modeled as a Markov Decision Process (MDP) to minimize the total cost. Each state consists of bilingual dictionaries and their completion status (dictionary size), an action set (machine and manual creations), a state transition probability (a likelihood that an action satisfies a dictionary size requirement), and a cost of action (time taken to complete state transition). The MDP outputs an optimal policy, which is a mapping from each state to its optimal action. To estimate total cost, costs of actions are accumulated by executing the actions with the highest state transition probabilities. The cost estimation is validated by an experiment described in the following chapter.

Chapter 6 presents a collaborative framework to conduct an experiment for validating a bilingual dictionary creation plan. The manual creation of bilingual dictionaries demands bilingual native speakers of low-resource languages. However, it is difficult to find such bilingual speakers. Therefore, this framework allows speakers of low-resource languages to collaborative in creating and evaluating bilingual dictionaries. An experiment is conducted on low-resource Indonesian languages with a minimum size threshold of 2,000 translation pairs. The result confirms the reliability of the proposed planning method: the actual total cost is 97% of the estimated total cost.

Chapter 7 concludes this thesis by summarizing the original contributions and future directions.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３,０００字を標準とすること。

論文内容の要旨を英語で記入する場合は、４００～１,１００words で作成し

審査結果の要旨は日本語５００～２,０００字程度で作成すること。

（続紙　２　）

（論文審査の結果の要旨）

本論文は、近縁言語間の対訳辞書を構築する手法の確立を目的としたものである。特に言語データの少ない低資源言語を対象とし、ピボット言語（中心軸となる言語）を用いた対訳辞書の自動生成手法を一般化し、さらにその生成時間を最適化する手法を提案している。得られた主要な成果は以下の通りである。

1.　言語間類似度に基づく近縁言語集合の抽出

対訳辞書の生成対象となる二言語（以下、対象言語）をピボット言語でつなぐことで、対象言語間の対訳辞書を高精度で生成する。一般に、こうした手法は、対象言語が類似している場合に効果的に働く。そこで、言語間類似度のデータベースを用いて、近縁言語のクラスタを同定する手法を提案している。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クラスタを抽出している。提案手法をインドネシア民族言語に実際に適用し、近縁言語のクラスタを抽出できることを確認している。

2.　ピボット言語に基づく対訳辞書生成のため制約最適化手法の一般化

従来から、対象言語とピボット言語間に予め存在する対訳辞書を用いて、対象二言語間の対訳辞書を新規に生成する手法が提案されてきた。しかし、既存の手法は、制約最適化アルゴリズムを用いて同根語のみを対訳として抽出するもので、得られた辞書の再現率が低い。そこで、本研究では従来手法を一般化し、制約を緩めることで、同根語以外の対訳を獲得している。具体的には、語義の共有度を用いて、同根語に加えてその同義語を対訳として抽出する。従来手法のF値は41%であったが、提案手法により64%にまで向上しており、その有効性が示されている。

3.　対訳辞書生成プランの最適化

低資源言語では、ピボット言語を用いた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。人手と機械の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作成をどのように組み合わせて手順化するかが問題となる。そこで、この問題をマルコフ決定問題として定式化し、所要時間を推定し、最善の対訳辞書生成手順を求める手法を提案している。提案手法をインドネシア民族言語に適用し評価実験を行った結果、実際の所要時間と推定所要時間の差は3%程度であり、提案手法の有効性が示されている。

以上、本論文は、近縁言語のクラスタを抽出し、近縁言語間の対訳辞書の自動生成手法と、対訳辞書生成手順の最適化手法を提案したもので、低資源言語間の対訳辞書構築に寄与するものである。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成30年8月9日に実施した論文内容とそれに関連した試問の結果、合格と認めた。

注）論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降