Indonesia Language Sphere: an ecosystem for dictionary development for low-resource languages

Yohei Murakami
Faculty of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, JAPAN
E-mail: yohei@fc.ritsumei.ac.jp

Abstract. There are more than 7000 languages around the world. However, 95% of the world population speak only 5% of them, at most 400 languages. More than half of them have fewer than 10,000 speakers. In 2010, UNESCO released a list of 2,464 endangered languages. In Indonesia, 144 languages are endangered. To preserve and increase the use of those languages, we started the Indonesia Language Sphere project. The purpose of this project is to develop comprehensive sets of bilingual dictionaries for Indonesian ethnic languages. To this end, we propose a generalized bilingual lexicon induction method that combines pairs of existing dictionaries. Furthermore, to reduce the total cost of bilingual dictionary creation, we combine the machine and manual creation processes and construct a planner that optimizes creation orders. This paper introduces the proposed methods and reports a preliminary experiment result focusing on Indonesian, Malay, Javanese, Sundanese, and Minangkabau.

1. Introduction
There are more than 7000 languages around the world, one-third of which are spoken in Asia [1]. The diversity of the languages in Asia is a major language barrier that prevents intercultural collaboration. Compared to Europe, languages of neighboring countries in Asia are not taught comprehensively and their scripts are totally different. To overcome the language barrier, we developed the multi-language service platform called the Language Grid [2]. We continue to operate it to provide various types of language services in different languages. Furthermore, because it was difficult for us to reach service providers in other countries, we started the federated operation of the Language Grid, where globally distributed grid operators operate their own local grids and exchange services with each other [3]. Initially, the federation was established among NECTEC in Thailand, University of Indonesia in Indonesia, and Xinjiang University in China. The Language Grid now has 183 participating groups from 24 countries sharing 226 language services.

However, the Language Grid currently fails to cover most Asian languages due to their resource scarcity, i.e., they have very scant electronic resources. According to LREMap [4] [5], a database storing the use and creation of language resources by collecting information during the submission process in NLP-related top conferences such as LREC, ACL, COLING, and so on, there are more than 5000 language resources in 214 languages and the only 11 languages have more than 100 resources. LREMap shows most of the languages are low-resource languages. To increase the resources for these languages, we have been working on the Indonesia
Language Sphere project since 2015\(^1\). The purpose of the project is to comprehensively cover Indonesia ethnic languages by creating many more bilingual dictionaries. Specifically, we create a new bilingual dictionary for two low resource languages by combining two existing bilingual dictionaries via a pivot language. We apply this method to many languages belonging to the same language family (hereafter called *intra-family languages*). As a result, the dictionaries depend on each other in an ecosystem where bilingual dictionaries are leveraged to establish a new one.

To this end, the project is addressing the following three issues.

**Pivot-base bilingual dictionary induction** A new bilingual dictionary is inductively generated from a graph connecting two bilingual dictionaries via a pivot language. Specifically, we formalize the identification of correct translation pairs as a weighted max SAT problem by introducing a semantic constraint based on language similarity. By solving this problem, we improve recall while matching the precision achieved by the existing inverse consultation method.

**Plan optimization of bilingual dictionary creation** To minimize the total cost of achieving comprehensive coverage of bilingual dictionaries for closely-related languages, it is critical to select the most appropriate pairs of languages. This decision involves a series of decisions, each with uncertainty in terms of state transition because the accuracy of dictionary induction and the size of the generated dictionary vary with language similarity and the size of the existing dictionaries. Therefore, we formalize the planning step as a Markov Decision Process in order to generate optimal plans.

The objective of this project is to develop comprehensive sets of bilingual dictionaries among closely related low-resource languages, especially Indonesian ethnic languages. To this end, we generalize 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, we combine machine and manual creations and optimize the order of dictionary creation. To sustainably develop, refine, and maintain the created dictionaries, we will deploy a bilingual dictionary service network on the Language Grid where each dictionary service interacts the others to propagate updated translation pairs.

The rest of this paper is organized as follows: In Section 2, we briefly introduce the background of the Language Grid. We explain our basic idea for creating bilingual dictionaries based on pivot languages using a constraint optimization technique in Section 3. Section 4 details the extension of the constraint-based approach to obtain many-to-many translation pairs. Section 5 introduces our plan optimization method to generate bilingual dictionary creation plans with minimal cost. Finally, Section 6 concludes our achievements on the Indonesia Language Sphere project.

### 2. Operation of the Language Grid

The Language Grid is a multi-language services platform for supporting intercultural collaboration. This platform allows users to share various language services and to combine them to create new language services customized for each user. We started operating an experimental infrastructure in 2007 to accumulate and share language resources as Web services. Note that our intention was not to gather language resources like ELRA Catalogue\(^2\) or LDC Catalog\(^3\), but rather to request that various institutions provide their language resources as services so that we could connect them to our servers.

---

1. http://langsphere.org/id/
2. http://catalogue.elra.info/en-us/
3. https://catalog.ldc.upenn.edu
In the 3 years since starting operation, 117 atomic language services became available on the Language Grid. The number of language services supporting Japanese and English was substantially higher than the other languages. These were followed by Chinese, Spanish, Korean, and Portuguese. These languages represent the barriers we often encounter in Japan because most language services were provided by Japanese institutions. We had difficulty in reaching service providers in other countries due to the unfamiliarity created by the separation. This emphasis on locality motivated us to launch a new service grid in other countries. To solve the bias of languages, we designed the federated operation of the Language Grid, where grid operators are globally dispersed, operate local grids, and exchange services with one another. Moreover, we extended our grid architecture to connect local grids to each other.

The federated operation of the Language Grid has allowed us to create a network of operation centers that cover various Asian languages. Operation centers were opened in Bangkok in 2010, Jakarta in 2011, and Urumqi in 2014; they have connected themselves to us to share a variety of services in Asian languages. As a result, 14 Asian WordNets are being shared through federation with Bangkok; Jakarta and Urumqi provide language services for Indonesian and Turkish language families.

Although the federation accelerated language service sharing and expanded the coverage of languages, difficulty remained in generating language services in low-resource languages. Based on the 5,758 entries listed in LREMap in 2016, 1999 resources relate to English (close to 34%, compared to 2% in 2012), followed by French (440), German (403), Spanish (294), Chinese (218) and Japanese (196), while there are still very few for Indonesia (13) and Malay (3) [4] [5]. Furthermore, Indonesia ethnic languages, even Javanese and Sundanese which have more than 30,000,000 speakers, have seen no resources submitted to the top conferences related to language resources. Figure 1 shows the statistics of language resources in LREMap according to language. The left vertical axis represents the number of language resources and the right vertical axis the ratio of a cumulative rate of speakers. There are 11 languages which have more than 100 resources and whose speakers occupy 54% of the world’s population. This means the remaining
speakers are not supported by adequate language resources. Therefore, we need technology to create language resources not limited to specific languages. We are actively developing bilingual dictionaries among closely related languages, especially Indonesia ethnic languages, 144 of which are endangered.

3. Pivot-based bilingual dictionary induction

To create a bilingual dictionary between intra-family languages, we employ a pivot-based method that induces a new bilingual dictionary by combining two existing bilingual dictionaries via a pivot language. However, this method can result in translation drift; the mistranslation of words due to pivot word polysemy. This drift prevents us from extracting correct translation pairs [6]. Wushouer et al. successfully selects correct translation pairs by formalizing the pivot-based method as a constraint optimization problem. This optimization problem employs a semantic constraint derived from the similarity of meaning of cognates between target languages. Specifically, they define a bilingual dictionary as a graph in which nodes are words and edges are translation relations. The two target language graphs are connected via pivot words in a common language among them, for example, Indonesia words in Fig. 2. They describe two constraints; first, all paths between both end nodes should be symmetric about the pivot language, and, second, translation pairs should be one to one pairs. Moreover, they model translation selection as a constraint optimization problem by allowing edges tagged with some costs to be added to asymmetric paths to satisfy the symmetry constraint [7].

Figure 2 shows an example of a graph that connects two bilingual dictionaries between Malay and Indonesian, and Indonesian and Minangkabau. Three reachable pairs, kutukan-kiparaik, sumpah-kiparaik, and sumpah-sumpah, are candidates of translation pairs. Firstly, kutukan-kiparaik is chosen because it satisfies the symmetry constraints without adding any edges that is no cost. Next, to satisfy the symmetry constraint between sumpah (Malay) and sumpah (Minangkabau), an edge is added between kutukan (Indonesian) and sumpah (Minangkabau) as shown in Fig. 2 and the reciprocal of the joint probability between sumpah (Malay) and sumpah (Minangkabau) is assigned to the edge as its cost. Specifically, the joint probability is approximately calculated by $P(\text{sumpah} (\text{Minangkabau})|\text{sumpah} (\text{Malay})) \times P(\text{sumpah} (\text{Minangkabau})|\text{sumpah} (\text{Malay})) = (1/2 \times 1) \times (1 \times 1) = 1/2$ and the cost is 2. Meanwhile, an edge can be added between sumpah (Indonesian) and kiparaik (Minangkabau). However, the cost is 4 because the joint probability between sumpah (Malay) and kiparaik (Minangkabau) is $(1/2 \times 1) \times (1 \times 1/2) = 1/4$. Since the latter cost is higher than the former, the sumpah (Malay) and sumpah (Minangkabau) is chosen as the second translation pair. Wushouer et al. add one more constraint called unique constraint constraint that means a translation of a word should be one word. Due to the unique constraint, the kutukan-kiparaik and sumpah-sumpah are extracted from Fig. 2.

Figure 2 shows an example of a graph that connects two bilingual dictionaries between
Malay and Indonesian, and Indonesian and Minangkabau. Three reachable pairs, kutukan-kiparaik, sumpah-kiparaik, and sumpah-sumpah, are candidates of translation pairs. First, kutukan-kiparaik is chosen because it satisfies the symmetry constraints without adding any edges, so there is no additional cost. Next, to satisfy the symmetry constraint between sumpah (Malay) and sumpah (Minangkabau), an edge is added between kutukan (Indonesian) and sumpah (Minangkabau) as shown in Fig. 2 and the reciprocal of the joint probability between sumpah (Malay) and sumpah (Minangkabau) is assigned to the edge as its cost. Specifically, the joint probability is approximately calculated by \( P(\text{sumpah}(\text{Minangkabau})|\text{sumpah}(\text{Malay})) \* P(\text{sumpah}(\text{Minangkabau})|\text{sumpah}(\text{Malay})) = (1/2 \* 1) \* (1 \* 1) = 1/2 \) and the cost is 2. Meanwhile, an edge can be added between sumpah (Indonesian) and kiparaik (Minangkabau). However, the cost is 4 because the joint probability between sumpah (Malay) and kiparaik (Minangkabau) is \((1/2 \* 1) \* (1 \* 1/2) = 1/4\). Since the latter cost is higher than the former, sumpah (Malay) and sumpah (Minangkabau) are chosen as the second translation pair. Wushouer et al. add one more constraint, called the unique constraint, that restricts the translation of a word to just one word. Due to the unique constraint, kutukan-kiparaik and sumpah-sumpah are extracted from Fig. 2.

4. Generalized constraint-based approach
The goal of Indonesia Language Sphere is to apply constraint-optimization-based bilingual dictionary induction to create bilingual dictionaries with full coverage for the languages used in Indonesia. However, different from Turkic languages that are assumed to be so close that each word in one language is mapped to one word in another language, Indonesian ethnic languages exhibit rather amorphous language similarities between them [8]. As a result, translation pairs should be many-to-many. Therefore, we generalize Wushouer’s method in order to acquire many-to-many translation pairs. To this end, we placed the unique constraint with a cost threshold. This threshold filters out incorrect translation pairs while preventing a decrease in the precision of the generated translation pairs. Furthermore, the existing bilingual dictionaries for Indonesia ethnic languages have lower quality and quantity than those of Turkic languages,
Table 1. Performance comparison between the generalized method and baselines.

| Case Study Method | Cognate Threshold | Precision | Recall | F-score |
|-------------------|-------------------|-----------|--------|---------|
| min-ind-zhm 2:T (M-M) | 6.0 | 0.713 | 0.953 | 0.815 |
| 1:T (M-M) | 4.6 | 0.836 | 0.713 | 0.770 |
| Baseline: 1:U (1-1) | 0.873 | 0.327 | 0.475 |
| Baseline: IC (M-M) | 0.950 | 0.031 | 0.059 |

\[\text{situatedMethod} ::= (\text{cycle}) : "(\text{constraint})\" \text{ where } \text{cycle}: \text{symmetry assumption cycle where cycle} \geq 1, \text{constraint}: \text{U as the unique constraint}[7], \text{and} \text{T as the threshold constraint}[9]. \text{IC: Inverse Consultation}[10]; 1-1 : one-to-one translation pair results; M-M : many-to-many translation pair results;\]

which are the official languages of each country. This means the existing dictionaries might lack essential translation pairs. To increase the recall of translation pairs by complementing the missing edges, we regard reachable pairs based on the added edges as candidates of translation pairs. Finally, we set up the threshold to optimize the F-score by using validation data. This proposal successfully generalizes constraint-optimization based bilingual dictionary induction to cover the case of bilingual dictionary creation between Turkic languages and can be applied to various intra-family languages [9].

We validated the proposed method by using it to create a Malay and Minangkabau bilingual dictionary from the source pairs of Indonesian and Malay, and Indonesian and Minangkabau. The language similarity between Malay and Minangkabau is 61.66%, which is smaller than that between Uyghur and Kazakh, 81.9%. As shown in Fig. 3, we identified the threshold that maximized the F-score of validation data as 6 and subsequent processing used this value. Table 1 summarizes the results. It shows that the proposed method (81.5% F-score) significantly outperformed the existing methods (47.5% F-score and 5.9% F-score). Moreover, the proposed method (2 cycle variant) has higher F-score than the 1 cycle variant. This confirms that the edge-adding cycle improves the performance, especially recall.

5. Plan optimization

To create comprehensive sets of bilingual dictionaries among multiple languages \((n \geq 3)\), we need to create \(n(n - 1)/2\) bilingual dictionaries. Between initial state (no dictionary) and final state (all the dictionaries), \(O(2^{n^2})\) states can exist. Furthermore, there can be many paths from the initial state to the final state by changing a pivot language of constraint-optimization-based method. The sheer size of this graph makes it difficult to find the optimal order of bilingual dictionary creation, i.e. that which minimizes the total cost. Moreover, if the languages are low-resource, there may not exist enough source bilingual dictionaries to perform machine creation. Manual creations of translation pairs are performed in that case. Therefore, we need to minimize the manual creation cost by optimizing the process of creating a set of bilingual dictionaries among closely related languages.

To optimally combine machine and manual processes, we model dictionary creation planning as a Markov Decision Process (MDP) targeting total cost minimization [11]. 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 rectangles in Fig. 4 represent states; the entries shown which language pairs have been created. A language pair with a dash means the size of the dictionary is not adequate. Circles represent actions: P means pivot action, whereas i means manual creation. Manual creation always transits to a state satisfying the user’s required size of the dictionary. On the other hand, the pivot action probabilistically transits to either satisfied state or unsatisfied state. MDP outputs an optimal
policy, which is a mapping from each state to its optimal action. The right of Fig. 4 shows the output of the Markov Decision Process for the case of 3 languages, the optimal policy is indicated by the blue circle. To estimate total cost, costs of actions are accumulated after executing the actions with the highest state transition probabilities.

We validated the proposed method by using it to create bilingual dictionaries for 5 languages. We focused on how much the cost could be decreased compared to all manual creation and how correctly it could estimate the total cost. The manual creation of bilingual dictionaries demands bilingual native speakers of low-resource languages. However, it is difficult to find such bilingual speakers. Therefore, we present a collaborative framework to conduct an experiment for validating the above bilingual dictionary creation plan. This framework allows speakers of low-resource languages to collaborate in creating and evaluating bilingual dictionaries. An experiment was conducted on low-resource Indonesian languages with a minimum size threshold of 2,000 translation pairs. The result confirmed the reliability of the proposed planning method: the actual total cost was 97% of the estimated total cost.

6. Conclusion

Indonesia has more than 700 languages. 144 languages of them are recognized as endangered languages. Moreover, even Indonesian itself is low-resource language. To save those languages, we are developing technologies to comprehensively create bilingual dictionaries among intra-family languages. Specifically, we have generalized a constraint-optimization-based method that uses a pivot language to generate many-to-many translation pairs and proposed a method that can output optimal plans in terms of minimizing the total cost. We have applied the proposals to Indonesian, Malay, Javanese, Sundanese, and Minangkabau to generate 2000 translation pairs for each bilingual dictionary. The cost was reduced by 42% compared to fully manual creation. Our technologies do not directly save endangered languages, but we hope these efforts will promote the usage of the Indonesia ethnic languages, and so remove them from the endangered list.

Acknowledgement

This research was partially supported by a Grant-in-Aid for Young Scientists (A) (17H04706, 2017-2020) from Japan Society for the Promotion of Science (JSPS).
References
[1] Paul L M, Simons G F and Fennig C D (eds.) 2015 Ethnologue: Languages of the World, Eighteenth edition (Dallas, Texas: SIL International)
[2] Ishida T, Murakami Y, Lin D, Nakaguchi T and Otani M 2018 Language Service Infrastructure on the Web: The Language Grid Computer 51(6) 72–81
[3] Murakami Y, Tanaka M, Lin D and Ishida T 2012 Service Grid Federation Architecture for Heterogeneous Domains International Conference on Services Computing (SCC 2012) (IEEE) pp 539–546
[4] Calzolari N, Gratta R D, Francopoulo G, Mariani J, Rubino F, Russo I and Soria C 2012 The LRE Map Harmonising Community Descriptions of Resources International Conference on Language Resources and Evaluation (LREC 2012) (ELRA) pp 1084–1089
[5] Gratta R.D, Frontini F, Khan A.F, Mariani J and Soria C 2014 The LREMap for Under-Resourced Languages International Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014) pp 78–83
[6] Tanaka R, Murakami Y and Ishida T 2009 Context-Based Approach for Pivot Translation Services International Joint Conference on Artificial Intelligence (IJCAI-09) pp 1555–61
[7] Wushouer M, Lin D, Ishida T and Hirayama K 2015 A constraint approach to pivot-based bilingual dictionary induction ACM Trans. Asian Low-Resour. Lang. Inf. Process. 15(1) 4:1–26
[8] Nasution A H, Murakami Y and Ishida T 2019 Generating Similarity Cluster of Indonesian Languages with Semi-Supervised Clustering International J. of Electrical and Computer Engineering (IJIECE) 9(1)
[9] Nasution A H, Murakami Y and Ishida T 2017 A generalized constraint approach to bilingual dictionary induction for low-resource language families ACM Trans. Asian Low-Resour. Lang. Inf. Process. 17(2) 9:1–29
[10] Tanaka K and Umemura K 1994 Construction of a Bilingual Dictionary Intermediated by a Third Language Proc. of the 15th Conference on Computational Linguistics (ACL-94) pp 297–303
[11] Nasution A H, Murakami Y and Ishida T 2017 Plan Optimization for Creating Bilingual Dictionaries of Low-Resource Languages International Conference on Culture and Computing (Kyoto: Culture and Computing) pp 35–41