Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages.

by Nasution Arbi Haza

Submission date: 30-Oct-2019 04:08PM (UTC+0800)
Submission ID: 1203421732
File name: 1.pdf (1.1M)
Word count: 5208
Character count: 28712

Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages

[1]Arbi Haza Nasution, [2]Yohei Murakami, [3]Toru Ishida
[1,3]Department of Social Informatics, Kyoto University, [2]Unit of Design, Kyoto University
Kyoto, Japan
[1]arbi@ai.soc.i.kyoto-u.ac.jp, [2]yohei@i.kyoto-u.ac.jp, [3]ishida@i.kyoto-u.ac.jp

Abstract

The constraint-based approach has been proven useful for inducing bilingual dictionary for closely-related low-resource languages. When we want to create multiple bilingual dictionaries linking several languages, we need to consider manual creation by a native speaker if there are no available machine-readable dictionaries are available as input. To overcome the difficulty in planning the creation of bilingual dictionaries, the consideration of various methods and costs, plan optimization is essential. Utilizing both constraint-based approach and plan optimizer, we design a collaborative process for creating 10 bilingual dictionaries from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We further design an online collaborative dictionary generation to bridge spatial gap between native speakers. We define a heuristic plan that only utilizes manual investment by the native speaker to evaluate our optimal plan with total cost as an evaluation metric. The optimal plan outperformed the heuristic plan with a 63.3% cost reduction.

Keywords: Bilingual Dictionary Creation, Low-resource Languages, Closely-related Languages

1. Introduction

Nowadays, machine-readable bilingual dictionaries are being utilized in actual services (Ishida, 2011) to support intercultural collaboration (Ishida, 2016; Nasution et al., 2017b), but low-resource languages lack such sources. Obviously bilingual lexicon extraction is highly problematic for low-resource languages due to the paucity or outright omission of parallel and comparable corpora. We introduced the promising approach of treating pivot-based bilingual dictionary induction for low-resource languages as an optimization problem (Nasution et al., 2016; Nasution et al., 2017c) where bilingual dictionaries are the only language resource required. Despite the high potential of our approach in enriching low-resource languages, it faces numerous issues when trying to create plans to implement multiple bilingual dictionaries for a set of low-resource languages like Indonesian ethnic languages. When actually implementing our constraint-based bilingual dictionary induction approach, we need to consider the inclusion of more traditional methods like manually creating the bilingual dictionaries by native speaker. In spite of the high cost, this will be unavoidable if no machine-readable dictionaries are available. Given the various methods and costs that may need to be considered, we recently introduced a plan optimizer to find the feasible optimal plan of creating multiple bilingual dictionaries with the least total cost (Nasution et al., 2017a). In this project, to create bilingual dictionary D_{A-B} between ethnic language L_A and ethnic language L_B, there is also a difficulty in finding a bilingual native speaker of two ethnic languages. To overcome this limitation, we can firstly create triple T_{A-ID-B} using the common language, Indonesian as pivot language L_{ID} where S_{ID-A}, a native bilingual speaker of Indonesian language L_{ID} - ethnic language L_A and S_{ID-B}, a native bilingual speaker of Indonesian language L_{ID} - ethnic language L_B collaborate by explaining the senses with Indonesian language. Then, the bilingual dictionary D_{A-B} can be induced from the triple T_{A-ID-B}. The native speakers need a tool that can bridge the spatial gap and help them collaborate. To actually implement our pivot-based bilingual dictionary induction following the optimal plan to create multiple Indonesian ethnic languages bilingual dictionaries, we address the following research goals:

- *Designing a Collaborative Process for Creating Bilingual Dictionaries of Indonesian Ethnic Languages*: Implementing plan optimization for creating bilingual dictionaries of low-resource languages and implementing a generalized constraint approach to bilingual dictionary induction for low-resource language families in creating 10 bilingual dictionaries with 2,000 translation pairs from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese.
- *Designing an Online Collaborative Dictionary Generation*: Bridging spatial gap between native speakers especially when doing a collaborative creation or evaluation of bilingual dictionary.

The rest of this paper is organized as follows: In Section 2 and Section 3, we will briefly discuss our constraint-based bilingual dictionary induction and plan optimizer, respectively. Section 4 details our collaborative process design. Finally, Section 5 concludes this paper.

2. Constraint-Based Bilingual Dictionary Induction

The traditional pivot-based approach is very suitable for low-resource languages (Tanaka and Umemura, 1994). Unfortunately, for some low-resource languages, it is often difficult to find machine-readable inverse dictionaries and corpora to identify and eliminate the erroneous translation

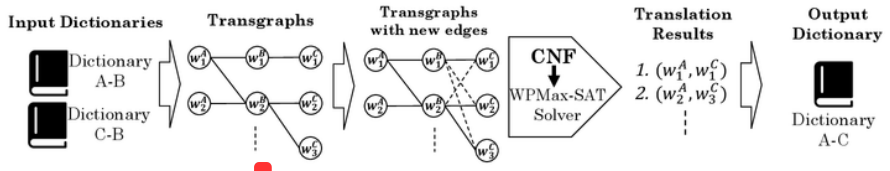


Figure 1: One-to-one constraint approach to pivot-based bilingual dictionary induction.

Figure 2: Modeling Bilingual Dictionary Induction Dependency.

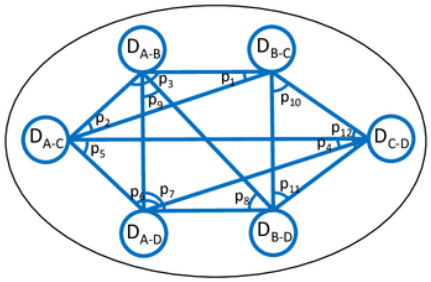


Figure 3: AND/OR Graph as an MDP State.

pair candidates. To overcome this limitation, our team (Wushouer et al., 2015) proposed to treat pivot-based bilingual lexicon induction as an optimization problem. The assumption was that lexicons of closely-related languages offer instances of one-to-one mapping and share a significant number of cognates (words with similar spelling/form and meaning originating from the same root language). The proposal uses a graph whose vertices represent words and edges indicate shared meanings; following (Soderland et al., 2009) it was called a transgraph. The proposal proceeds as follows: (1) use two bilingual dictionaries as input, (2) represent them as transgraphs where w_1^A and w_2^A are non-pivot words in language L_A, w_1^B and w_2^B are pivot words in language L_B, and w_1^C, w_2^C and w_3^C are non-pivot words in language L_C, (3) add some new edges represented by dashed edges based on the one-to-one assumption, (4) formalize the problem into conjunctive normal form (CNF) and use the Weighted Partial MaxSAT (WPMaxSAT) solver (Ansótegui et al., 2009) to return the optimized translation results, and (5) output the induced bilingual dictionary as the result. These steps are shown in Figure 1. However, the assumption of one-to-one mapping is too strong to induce the many-to-many translation pairs needed to offset resource paucity because few such pairs can be found. Therefore, we generalized the constraint-based bilingual dictionary induction framework by extending constraints and translation pair candidates from the one-to-one approach to attain more voluminous bilingual dictionary results with many-to-many translation pairs extracted from connected existing and new edges (Nasution et al., 2016). We further enhance our generalized method by setting two steps to obtaining translation pair results. First, we identify one-to-one cognates by incorporating more constraints and heuristics to improve the quality of the translation result. We then identify the cognates' synonyms to obtain many-to-many translation pairs. In each step, we can obtain more cognate and cognate synonym pair candidates by iterating the n-cycle symmetry assumption until all possible translation pair candidates have been reached (Nasution et al., 2017c).

3. Plan Optimizer

Our constraint-based bilingual dictionary induction approach has the potential to enrich low-resource languages with the only input being machine readable bilingual dictionaries. Unfortunately, the scarcity of such dictionaries for low-resource languages makes it difficult to plan which bilingual dictionary should be invested first or which bilingual dictionary should be induced right from the start in order to obtain all possible combination of bilingual dictionaries from the language set with the minimum total cost to be paid. We model the bilingual dictionary dependency with AND/OR graphs as shown in Figure 2, and employ the Markov Decision Process (MDP) for plan optimization where a state is defined by AND/OR graphs as shown in Figure 3. The exponential complexity of formulating the bilingual dictionary creation planning into a graph theory problem indicates a greater complexity of obtaining the optimal planning with the least total cost by only following the heuristic. Nevertheless, our algorithm greatly reduced the complexity, so that the MDP planning can find the feasible optimal plan with less total cost compared to heuristic planning (e.g., only use manual investment by native speaker). Our MDP model can calculate the cumulative cost while predicting and considering the probability of the pivot action to yield a satisfying output bilingual dictionary as utility for every state to better predict the most feasible optimal plan with the least total cost. Our formalization with MDP allow user to predict the feasible optimal plan with the least total cost before implementing the constraint-based bilingual dictionary induction framework in a big scale.

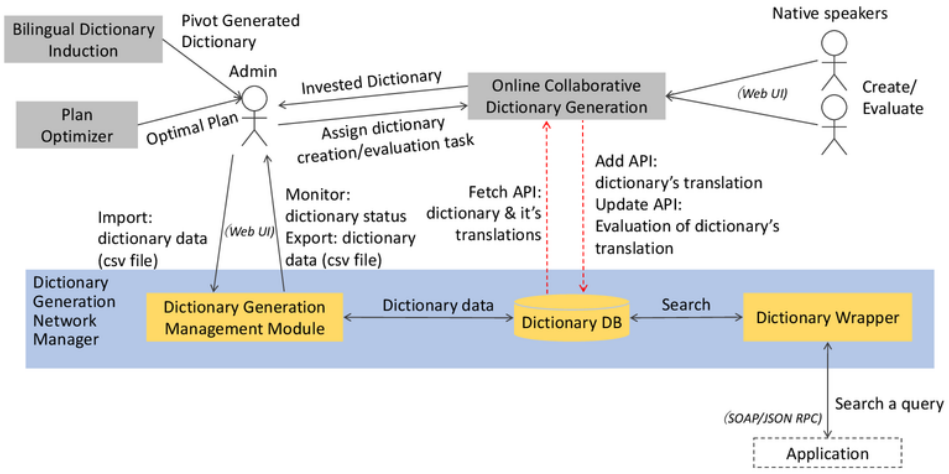


Figure 4: System Integration Overview.

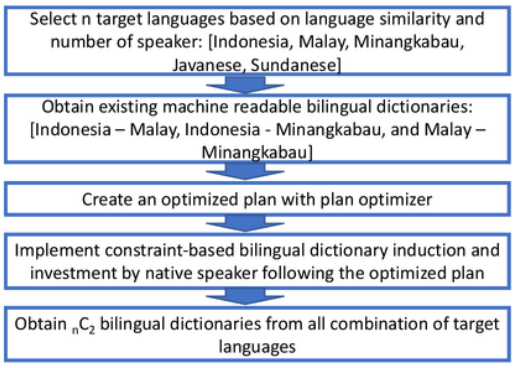


Figure 5: Overview of Bilingual Dictionaries Generation Process.

4. Designing a Collaborative Process

4.1. Overview

We integrate our Constraint-based Bilingual Dictionary Induction and Plan Optimizer with an Online Collaborative Dictionary Generation as a tool to bridge the spacial gap between native speakers and a Dictionary Generation Network Manager to manage the final dictionary so that it is accessible via API in the Language Grid (Ishida, 2011) as shown in Figure 4. The overview of bilingual dictionaries generation process is shown in Figure 5 while the detailed process is explained in Algorithm 1.

4.2. Selecting Target Languages

To select target languages in this paper, we use an Automatic Similarity Judgment Program (ASJP) (Holman et al., 2011) following our previous work (Nasution et al., 2017d). Indonesia has 707 low-resource ethnic languages (Lewis et al., 2015) that require our attention. There are two factors we consider in selecting the target languages: language similarity and number of speakers. In order to ensure that the induced bilingual dictionaries will be useful for many users, we listed the top 10 Indonesian ethnic languages ranked by the number of speakers. Since our constraint-based approach works better on closely related languages, we further generated the language similarity matrix by utilizing ASJP as shown in Table 1. Based on number of speaker, we select Javanese and Sundanese. To find and coordinate native speakers of those languages, we collaborate with Telkom University. Based on relatedness with Indonesian, we select Malay and Minangkabau. To find and coordinate native speakers of those language, we collaborate with Islamic University of Riau. Hence, we target 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We want to enrich/create the following dictionaries: Indonesia-Malay, Indonesia-Minangkabau, Indonesia-Javanese, Indonesia-Sundanese, Malay-Minangkabau, Malay-Javanese, Malay-Sundanese, Minangkabau-Javanese, Minangkabau-Sundanese, and Javanese-Sundanese with 2,000 translation pairs each.

4.3. Modeling Task for Native Speaker

When actually implementing our constraint-based bilingual dictionary induction approach, we need native speakers for manual creation of bilingual dictionaries or evaluation of the output dictionaries. There are a lot of prior researches on modeling workflow management (Georgakopoulos et al., 1995; Hollingsworth and Hampshire, 1995; Kappel et al., 2000; Huang et al., 2000; Alexopoulos et al., 2011; Kulkarni et al., 2012). We define several rules of which native speaker can create/evaluate which dictionary.

A bilingual dictionary between ethnic language L_A and ethnic language L_B, D_{A-B} can be induced from a triple T_{A-ID-B}, while a triple T_{A-ID-B} can be induced from a bilingual dictionary D_{ID-A} and a bilingual dictionary D_{ID-B}. A bilingual dictionary between Indonesian language L_{ID} and ethnic language L_A, D_{ID-A} can be man-

Algorithm 1: Bilingual Dictionaries Generation

Input: targetLanguageInfo, existingDictionaries
/* In this project, targetLanguages: [Indonesia,Malay,Minangkabau,Javanese,Sundanese] */
/* targetLanguageInfo includes language similarities and expectedDictionarySize=2,000 */
/* existingDictionaries=[$D_{Indonesia-Malay}$, $D_{Indonesia-Minangkabau}$, $D_{Malay-Minangkabau}$] */
Output: dictionaryList /* all combination of bilingual dictionaries from the targetLanguages */

for *each* D_{A-B} *in existingDictionaries* **do**
  dictionaryList.add(D_{A-B});
end
optimizedPlan ← planOptimizer.create(targetLanguageInfo, dictionaryList);
for *each action to create bilingual dictionary* D_{A-B} *in optimizedPlan* **do**
  if *final state is reached* **then**
    return dictionaryList
  end
  else
    if *action type = investment* **then**
      /* CT1(L_{ID}, L_A): Creation and Evaluation of Indonesia-Ethnic Bilingual Dict */
      if L_A *or* L_B *is Indonesian language* L_{ID} **then**
        create and evaluate bilingual dictionary D_{A-B} by a native bilingual speaker S_{A-B};
        dictionaryList.add(D_{A-B});
      end
      /* CT2(L_A, L_B): Creation and Evaluation of Ethnic-Ethnic Bilingual Dict */
      else
        if *native bilingual speaker* S_{A-B} *is available* **then**
          create and evaluate bilingual dictionary D_{A-B} by a native bilingual speaker S_{A-B};
          dictionaryList.add(D_{A-B});
        end
        else
          create and evaluate triple T_{A-ID-B} by two native bilingual speakers S_{ID-A} and S_{ID-B};
          induce D_{A-B} from T_{A-ID-B}; dictionaryList.add(D_{A-B});
        end
      end
    end
    else if *action type = pivot* **then**
      use constraint-based approach to obtain triple T_{A-P-B} ;
      /* T4 (L_A, L_P, L_B) */
      if *native bilingual speaker* S_{A-B} *is available* **then**
        evaluate triple T_{A-P-B} by a native bilingual speaker S_{A-B};
        induce D_{A-B} from T_{A-P-B}; dictionaryList.add(D_{A-B});
      end
      else
        evaluate triple T_{A-P-B} by two native bilingual speakers S_{ID-A} and S_{ID-B};
        induce D_{A-B} from T_{A-P-B}; dictionaryList.add(D_{A-B});
      end
    end
  end
end

ually created or evaluated by a native bilingual speaker S_{ID-A}. A bilingual dictionary D_{A-B} can be manually created or evaluated by a native bilingual speaker S_{ID-A} and a native bilingual speaker S_{ID-B} collaboratively or by a native bilingual speaker S_{A-B} alone.

There are some bilingual dictionaries between Indonesian and Indonesian ethnic languages exist in a printed format. We may be able to digitalized the printed Indonesian - ethnic language bilingual dictionaries to a machine readable format. Nevertheless, when we connect the digitalized bilingual dictionary D_{ID-A} and a bilingual dictionary D_{ID-B} via Indonesian language L_{ID} as a pivot, and further induced D_{A-B} with our constraint-based approach, we expect that there will be many unreachable translation pair candidates since some Indonesian words in one bilingual dictionary may not exist in the other bilingual dictionary. In order to maximize the use of our pivot-based approach, we prepare a list of 2,000 most commonly used Indonesian words to be translated to ethnic language L_A to create a bilingual dictionary D_{ID-A} by a native bilingual speaker S_{ID-A} as shown in Figure 6. Due to budget limitation, we only allow the native speaker to translate an Indonesian word to up to five words of ethnic language L_A. To ensure the quality of the manually created bilingual dictionary D_{ID-A}, another native bilingual speaker S_{ID-A} will evaluate the translation pairs as shown

Table 1: Similarity Matrix of Top 10 Indonesian Ethnic Languages Ranked by Number of Speakers

Language	**Indonesian**	Malang	Yogyakarta	**Javanese**	**Sundanese**	**Malay**	Palembang Malay	Madurese	**Minangkabau**
Malang	23.46%								
Yogyakarta	27.29%	87.36%							
Javanese	**24.09%**	47.50%	52.18%						
Sundanese	**39.43%**	18.55%	22.43%	**21.82%**					
Malay	**85.10%**	20.53%	24.35%	**21.36%**	**41.12%**				
Palembang Malay	68.24%	33.97%	37.97%	31.85%	38.90%	73.23%			
Madurese	34.45%	17.63%	14.15%	15.18%	19.86%	34.16%	34.32%		
Minangkabau	**61.59%**	26.59%	29.63%	**25.01%**	**30.81%**	**61.66%**	63.60%	34.32%	
Buginese	31.21%	12.76%	16.85%	18.33%	24.80%	32.04%	31.00%	17.94%	32.00%

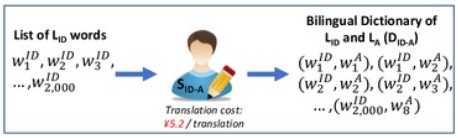


Figure 6: $T1(L_{ID}, L_A)$: Creation of Bilingual Dictionary D_{ID-A}.

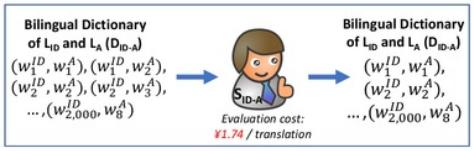


Figure 7: $T2(L_{ID}, L_A)$: Evaluation of Bilingual Dictionary D_{ID-A}.

in Figure 7. To overcome the limitation in finding native bilingual speakers of two ethnic languages for creation and evaluation of bilingual dictionary D_{A-B}, two native bilingual speakers S_{ID-A} and S_{ID-B} can collaborate as shown in Figure 8 and Figure 9 respectively. Finally, there are two composite tasks, which are $CT1(L_{ID}, L_A)$, a manual creation followed by evaluation of bilingual dictionary D_{ID-A} as shown in Figure 10a and $CT2(L_A, L_{ID}, L_B)$, a manual creation followed by evaluation of bilingual dictionary D_{A-B} as shown in Figure 10b.

4.4. Online Collaborative Dictionary Generation

The online collaborative dictionary generation has 6 modules: individual creation of Indonesia-Ethnic bilingual dictionary, individual evaluation of Indonesia-ethnic bilingual dictionary, individual creation of ethnic-ethnic bilingual dictionary, individual evaluation of ethnic-ethnic bilingual dictionary, collaborative creation of ethnic-ethnic bilingual dictionary, and collaborative evaluation of ethnic-ethnic bilingual dictionary. Each native speakers get his/her own user account. They can login to the system, read the user manual, update their profile, check their assigned task, and do their assigned task. For the individual task, the native speakers can do the task anywhere before the deadline as shown in Figure 11. However, for the collaborative task, a pair of native speakers need to login to the system at the same time in order to collaborate. The live chat is used to

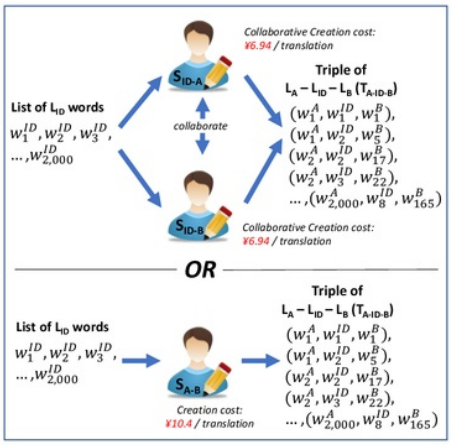


Figure 8: $T3(L_A, L_{ID}, L_B)$: (Individual/Collaborative) Creation of Triple T_{A-ID-B} to induce Bilingual Dictionary D_{A-B}.

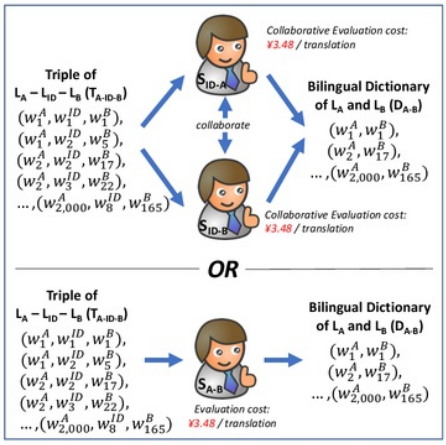


Figure 9: $T4(L_A, L_{ID}, L_B)$: (Individual/Collaborative) Evaluation of Triple T_{A-ID-B} to induce Bilingual Dictionary D_{A-B}.

ease communication and discussion during the collaborative creation / evaluation session as shown in Figure 12.

Table 2: Estimated Cost of Actions following MDP Optimal Plan

Task following MDP Plan	#Translation[1]	MDP Transition Probability[2]	Estimated Precision[2]	Unit Cost (JPY)	Total Cost (JPY)
T1(Indonesian, Malay)	1,480[3]			5.20	7,696.00
T2(Indonesian, Malay)	1,480			1.74	2,575.00
T1(Indonesian, Javanese)	2,000			5.20	10,400.00
T2(Indonesian, Javanese)	2,000			1.74	3,480.00
T1(Indonesian, Sundanese)	2,000			5.20	10,400.00
T2(Indonesian, Sundanese)	2,000			1.74	3,480.00
P(Malay, Indonesia, Minangkabau)	1,645[3]	0.983	0.4113	0.00	0.00
T4(Malay, Indonesian, Minangkabau)	754			6.96	5,248.00
P(Javanese, Indonesian, Sundanese)	1,027	0.972	0.2567	0.00	0.00
T4(Javanese, Indonesian, Sundanese)	1,027			6.96	7,147.00
T3(Javanese, Sundanese)	973			13.88	13,507.00
T4(Javanese, Sundanese)	973			6.96	6,773.00
P(Malay, Indonesia, Javanese)	1,094	0.943	0.2481	0.00	0.00
T4(Malay, Indonesia, Javanese)	1,094			6.96	7,615.00
T3(Malay, Javanese)	906			13.88	12,575.00
T4(Malay, Javanese)	906			6.96	6,305.00
P(Minangkabau, Indonesian, Sundanese)	1,157	0.949	0.289	0.00	0.00
T4(Minangkabau, Indonesian, Sundanese)	1,157			6.96	8,049.00
T3(Minangkabau, Sundanese)	844			13.88	11,708.00
T4(Minangkabau, Sundanese)	844			6.96	5,871.00
P(Malay, Indonesian, Sundanese)	1,356	0.826	0.3045	0.00	0.00
T4(Malay, Indonesian, Sundanese)	1,356			6.96	9,434.00
T3(Malay, Sundanese)	645			13.88	8,946.00
T4(Malay, Sundanese)	645			6.96	4,486.00
P(Minangkabau, Malay, Javanese)	1,148	0.929	0.2608	0.00	0.00
T4(Minangkabau, Malay, Javanese)	1,148			6.96	7,993.00
T3(Minangkabau, Javanese)	852			13.88	11,820.00
T4(Minangkabau, Javanese)	852			6.96	5,927.00
TOTAL					**171,435.00**

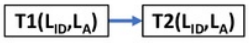
[1] *A number of translations is calculated from the number of translation pair candidates from the constraint-based approach × estimated precision with a high polysemy rate.*
[2] *Estimated from beta distribution based on language similarity and high polysemy pivot rate following our unpublished ACM TALLIP article entitled "Plan Optimization to Bilingual Dictionary Induction for Low-Resource Language Families".*
[3] *Excluding translation pairs from existing bilingual dictionaries: Indonesian-Malay (520 translation pairs) and Malay-Minangkabau (1,246 translation pairs).*

table

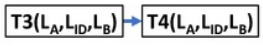
Table 3: Estimated Cost of Actions following Heuristic Plan

Task following Heuristic Plan	#Translation[1]	Unit Cost (JPY)	Total Cost (JPY)
T1(Indonesian, Javanese)	2,000	5.20	10,400.00
T2(Indonesian, Javanese)	2,000	1.74	3,480.00
T1(Indonesian, Sundanese)	2,000	5.20	10,400.00
T2(Indonesian, Sundanese)	2,000	1.74	3,480.00
T1(Indonesian, Malay)	1,480[1]	5.20	7,696.00
T2(Indonesian, Malay)	1,480	1.74	2,575.20
T3(Javanese, Sundanese)	2,000	13.88	27,760.00
T4(Javanese, Sundanese)	2,000	6.96	13,920.00
T3(Malay, Minangkabau)	754[1]	13.88	10,465.52
T4(Malay, Minangkabau)	2,000	6.96	13,920.00
T3(Malay, Javanese)	2,000	13.88	27,760.00
T4(Malay, Javanese)	2,000	6.96	13,920.00
T3(Minangkabau, Sundanese)	2,000	13.88	27,760.00
T4(Minangkabau, Sundanese)	2,000	6.96	13,920.00
T3(Malay, Sundanese)	2,000	13.88	27,760.00
T4(Malay, Sundanese)	2,000	6.96	13,920.00
T3(Minangkabau, Javanese)	2,000	13.88	27,760.00
T4(Minangkabau, Javanese)	2,000	6.96	13,920.00
TOTAL			**270,816.72**

[1] *Excluding translation pairs from existing bilingual dictionaries: Indonesian-Malay (520 translation pairs) and Malay-Minangkabau (1,246 translation pairs).*



(a) $CT1(L_{ID}, L_A)$: Composite Task Creation and Evaluation of Bilingual Dictionary D_{ID-A}.



(b) $CT2(L_A, L_{ID}, L_B)$: Composite Task Creation and Evaluation of Bilingual Dictionary D_{A-B}.

Figure 10: Composite Tasks.

4.5. Cost Estimation

We estimate the cost of each native speaker tasks as follows:

- $T1(L_{ID}, L_A)$: From an estimated duration of 30 seconds per translation and a daily wage of JPY5,000/8 hours, the estimated total translation per day is $1 \times 2 \times 60 \times 8 = 960$ and the estimated cost is JPY5.2 per correct translation.
- $T2(L_{ID}, L_A)$: From an estimated duration of 10 seconds per translation and a daily wage of JPY5,000/8 hours, the estimated total translation per day is $1 \times 6 \times 60 \times 8 = 2,880$ and the estimated cost is JPY1.74 per

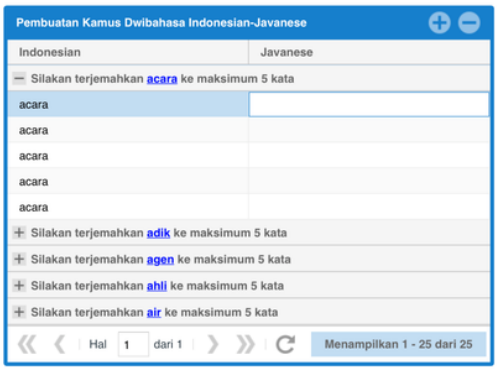


Figure 11: Individual Creation of Indonesia-Ethnic Bilingual Dictionary.

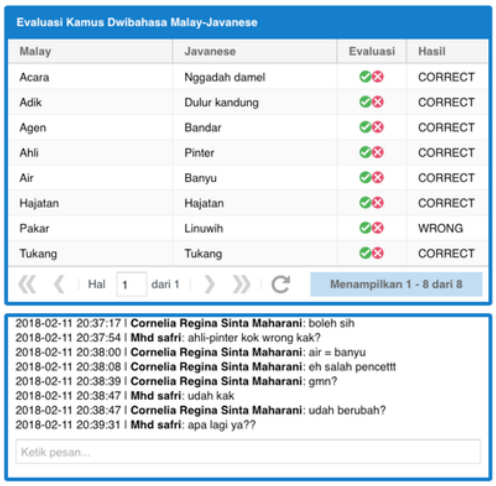


Figure 12: Collaborative Evaluation of Ethnic-Ethnic Bilingual Dictionary.

correct translation.

- $T3(L_A, L_{ID}, L_B)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, for the individual task, from an estimated duration of 60 seconds per translation, the estimated cost is JPY5.2×2 = JPY10.4 per translation. For the collaborative task, from an estimated duration of 30 seconds to translate an Indonesian word to each ethnic language in parallel, and an extra 10 seconds for discussing the sense sharing between the two ethnic language translations, the estimated total cost is (JPY5.2 + JPY11.74)×2 workers = JPY13.88 per correct translation pair.

- $T4(L_A, L_{ID}, L_B)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, for the individual task, from an estimated duration of 20 seconds per translation, the estimated cost is JPY1.74×2 = JPY3.48 per translation. For the collaborative task, from an estimated duration of 20 seconds to evaluate by discussing the sense sharing between the two ethnic language translations, the estimated total cost is (JPY1.74 + JPY1.74)×2 workers = JPY6.96 per correct translation pair.

- $CT1(L_{ID}, L_A)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, the estimated cost is JPY5.2 + JPY1.74 = JPY6.94 per translation.

- $CT2(L_A, L_B)$: Following the cost of $T3(L_A, L_{ID}, L_B)$ and $T4(L_A, L_{ID}, L_B)$ and the combination of workers based on availability of native bilingual speakers ($S_{A-B} + S_{A-B}$, $S_{A-B} + S_{ID-A}\&S_{ID-B}$, $S_{ID-A}\&S_{ID-B} + S_{A-B}$, $S_{ID-A}\&S_{ID-B} + S_{ID-A}\&S_{ID-B}$), the variations of estimated total cost are (JPY10.4 + JPY3.48 = JPY13.88, JPY10.4 + JPY6.96 = JPY17.36, JPY13.88 + JPY3.48 = JPY17.36, JPY13.88 + JPY6.96 = JPY20.84) respectively.

We estimate the cost of actions following the optimized plan utilizing both constraint-based approach and manual investment by native speakers as shown in Table 2 and further compare them with cost of actions following the heuristic plan utilizing only manual investment by native speakers as shown in Table 3.

5. Conclusion

We design a collaborative process for creating 10 bilingual dictionaries with 2,000 translation pairs from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We implement our plan optimizer and our generalized constraint approach to bilingual dictionary induction in creating input dictionaries or evaluating the resulting bilingual dictionaries. We define a heuristic plan that only utilize manual investment by native speaker to evaluate our optimal plan with total cost as an evaluation metric. By following the optimal plan, we can reduce 63.3% cost of following the heuristic plan. We further design an online dictionary generation tool to bridge spatial gap between native speakers. We will analyze the native speakers' behavior and chat log for future improvement of the system.

6. Acknowledgment

This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020) and a Grant-in-Aid for Young Scientists (A) (17H04706, 2017-2020) from Japan Society for the Promotion of Science (JSPS). The first author was supported by Indonesia Endowment Fund for Education (LPDP). We thank our collaborators, Department of Informatics Engineering, Universitas Islam Riau and Department of Computational Science, Telkom University for their effort in gathering the student workers.

7. Bibliographical References

Alexopoulos, K., Makris, S., Xanthakis, V., and Chryssolouris, G. (2011). A web-services oriented workflow

management system for integrated digital production engineering. *CIRP Journal of Manufacturing Science and Technology*, 4(3):290 – 295. Production Networks Sustainability.

Ansótegui, C., Bonet, M. L., and Levy, J. (2009). Solving (weighted) partial maxsat through satisfiability testing. In *Theory and Applications of Satisfiability Testing-SAT 2009*, pages 427–440. Springer.

Georgakopoulos, D., Hornick, M., and Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2):119–153, Apr.

Hollingsworth, D. and Hampshire, U. (1995). Workflow management coalition: The workflow reference model. *Document Number TC00-1003*, 19.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., et al. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Huang, G., Huang, J., and Mak, K. (2000). Agent-based workflow management in collaborative product development on the internet. *Computer-Aided Design*, 32(2):133 – 144.

Toru Ishida, editor. (2011). *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer Publishing Company, Incorporated.

Ishida, T. (2016). Intercultural collaboration and support systems: A brief history. In *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2016)*, pages 3–19. Springer.

Kappel, G., Rausch-Schott, S., and Retschitzegger, W. (2000). A framework for workflow management systems based on objects, rules and roles. *ACM Comput. Surv.*, 32(1es), March.

Kulkarni, A., Can, M., and Hartmann, B. (2012). Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1003–1012, New York, NY, USA. ACM.

M. Paul Lewis, et al., editors. (2015). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 18th edition.

Nasution, A. H., Murakami, Y., and Ishida, T. (2016). Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3291–3298, Paris, France, May.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017a). Plan optimization for creating bilingual dictionaries of low-resource languages. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 35–41, Sept.

Nasution, A. H., Syafitri, N., Setiawan, P. R., and Suryani, D. (2017b). Pivot-based hybrid machine translation to support multilingual communication. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 147–148, Sept.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017c). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):9:1–9:29, November.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017d). Similarity cluster of indonesian ethnic languages. In *Proceedings of the First International Conference on Science Engineering and Technology (ICoSET 2017)*, pages 12–27, Pekanbaru, Indonesia, November.

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., et al. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 262–270. Association for Computational Linguistics.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

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):4:1–4:26, November.

Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages.

ORIGINALITY REPORT

18%	17%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ link.springer.com

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		