2017 International Conference on Culture and Computing

10-12 September 2017 • Kyoto, Japan

CONFERENCE INFORMATION

PAPERS BY SESSION

PAPERS BY AUTHOR

GETTING STARTED

TRADEMARKS

SEARCH

Proceedings

2017 International Conference on Culture and Computing

Culture and Computing 2017

Kyoto, Japan
10-12 September 2017

Proceedings

2017 International Conference on Culture and Computing

Culture and Computing 2017

Kyoto, Japan
10-12 September 2017

CPS
Conference Publishing Services

IEEE Computer Society Order Number E6288
ISBN-13: 978-1-5386-1135-7
BMS Part # CFP1710R-CDR

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Juan E. Guerrero
Cover art production by Annie Jiu
Printed in the United States of America by Applied Digital Imaging

IEEE Computer Society
***Conference Publishing Services* (CPS)**
http://www.computer.org/cps

2017 International Conference on Culture and Computing

Culture and Computing 2017

Table of Contents

Session 1: Cultural Heritage and Archiving

Session 2: Language and Infrastructure for Culture

Organized Session: Intangible Cultural Properties

Session 3: Human Behaviors and Culture

Session 4: Information Technologies for Performing Arts

Demo Session & Exhibition

Session 5: Media Art and Culture

Poster Session

Session 6: Music and Culture

Message from the Chairs
Culture and Computing 2017

Welcome to 2017 International Conference on Culture and Computing (Culture and Computing 2017).

International communities are facing various problems in different topic areas such as: population demographic shifts, energy use and creation, the environment, and food supply. It is necessary to build a global consensus for resolving problems within these topic areas. Unfortunately, there are difficulties that hinder communication among cultures. It is imperative to develop information and communication technologies that encourage mutual understanding and bridge the difference in cultures.

Several research directions impinge on the relations between culture and computing: archiving cultural heritage via ICT (cf. digital archives), empowering humanities researches via ICT (cf. digital humanities), creating art and expressions via ICT (cf. media art), generating culturally-directed behavior (cf. cultural agent), supporting multi-language, multi-cultural societies via ICT (cf. intercultural collaboration), and understanding new cultures born in the Internet and the Web (cf. net culture).

This year, Culture and Computing 2017 is held in Kyoto, the cultural heart of Japan, to provide an opportunity to share research issues and discuss the future of culture and computing. The conference proceedings include full papers and short papers for oral presentation, poster papers and demonstration papers. The oral presentation track will present a collection of scientific or engineering research results that include six regular sessions and one organized session: cultural heritage and archiving, language and infrastructure for culture, human behaviors and culture, information technologies for performing arts, media art and culture, music and culture, and intangible cultural properties. The poster session and the demonstration session will provide thought-provoking stimulation and discussion. We are sure you will find your participation in the conference fruitful and hope that it is enjoyable.

We wish to express our sincere appreciation to everyone who has contributed to Culture and Computing 2017. Our foremost thanks go to the authors of the submitted papers from 11 countries for their valuable ideas and substantial efforts. We are also very grateful for the hard work of the program committee members and external reviewers who have helped to ensure the quality of the Culture and Computing program. Without them, the review process would not have been so thorough and effective. Finally, we wish to thank the IEEE Computer Society Conference Publishing Services for their support in compiling the proceedings.

We hope you enjoy the stimulating Culture and Computing program and your stay in Kyoto!

Toru Ishida, Naoko Tosa, and Kozaburo Hachimura
Culture and Computing 2017 General Co-chairs

Mamiko Sakata, Donghui Lin, Akira Maeda
Culture and Computing 2017 Program Co-chairs

Conference Organization

Culture and Computing 2017

Honorary Chair

Jin Mingzhe, *Doshisha University, Japan*

General Co-Chairs

Kozaburo Hachimura, *Ritsumeikan University, Japan*
Toru Ishida, *Kyoto University, Japan*
Naoko Tosa, *Kyoto University, Japan*

Program Co-Chairs

Mamiko Sakata, *Doshisha University, Japan*
Donghui Lin, *Kyoto University, Japan*
Akira Maeda, *Ritsumeikan University, Japan*

Local Arrangement Co-Chairs

Akihiro Kawase, *Doshisha University, Japan*
Kenji Hatano, *Doshisha University, Japan*

Local Arrangement Committee

Yanyan Chen, *Doshisha University, Japan*
Tamaki Yano, *Doshisha University, Japan*
Noriko Suzuki, *Tezukayama University, Japan*
Satoru Tanaka, *Ritsumeikan University, Japan*

Poster Co-Chairs

Michiya Yamamoto, *Kwansei Gakuin University, Japan*
Saizo Aoyagi, *Toyo University, Japan*

Art Chair

Kanako Hamasaki, *Kodokan, Japan*

Publicity Chair

Gen Tsuchiyama, *Doshisha University, Japan*

Asian Liaison

Ryohei Nakatsu, *Kyoto University, Japan*

North American Liaison

Katsushi Ikeuchi, *Microsoft Corp., USA*

Program Committee

Culture and Computing 2017

Francisco Grimaldo Moreno, *University of Valencia, Spain*
Jens Allwood, *IDÛ Interactive Inc, Sweden*
Timothy Bickmore, *Northeastern University – Boston, USA*
Emmanuel G. Blanchard, *IDÛ Interactive Inc., Canada*
Jean-Pierre Briot, *CNRS, Brazil*
Anna Carbone, *Polytechnic University of Turin, Italy*
Philippe Codognet, *CNRS & University of Tokyo, Japan*
Nick Degens, *Wageningen University, Netherlands*
Christiane Fellbaum, *Princeton University, USA*
Rüdiger Heimgärtner, *IUIC – Undorf, Germany*
Michitaka Hirose, *University of Tokyo, Japan*
Gert Jan Hofstede, *Wageningen University, Netherlands*
Jieh Hsiang, *National Taiwan University, Taiwan*
Katsushi Ikeuchi, *Microsoft Corp., USA*
Jean Ippolito, *University of Hawaii at Hilo, USA*
Lewis Johnson, *Alelo Inc. USA*
Yasuhiro Katagiri, *Future University Hakodate, Japan*
Tomoko Koda, *Osaka Institue of Technology, Japan*
Sadao Kurohashi, *Kyoto University, Japan*
Lydia Lau, *University of Leeds, United Kingdom*
Michihiko Minoh, *Kyoto University, Japan*
Shigeru Miyagawa, *MIT, USA*
Shigeo Morishima, *Waseda University, Japan*
Yohei Murakami, *Kyoto University, Japan*
Ryohei Nakatsu, *Kyoto University, Japan*
Atsushi Nakazawa, *Kyoto University, Japan*
Shohei Nobuhara, *Kyoto University, Japan*
Hiroaki Ohshima, *University of Hyogo, Japan*
Masayuki Ohtani, *Kindai University, Japan*
Yoshihiro Okada, *Ryukoku University, Japan*
Masashi Okubo, *Doshisha University, Japan*
Mario Paolucci, *ISTC-CNR, Italy*
Jong-Il Park, *Hanyang University, Korea*
Andrew Prescott, *University of Glasgow, United Kingdom*
Walter Quattrociocchi, *IMT School for Advanced Studies Lucca, Italy*
Matthias Rehm, *Aalborg University, Denmark*
Geoffrey Rockwell, *University of Alberta, Canada*
Kasper Rodil, *Aalborg University, Denmark*
Tetsuo Sawaragi, *Kyoto University, Japan*
Vibeke Sorensen, *Nanyang Technological University, Singapore*
Virach Sornlertlamvanich, *Thammasat University, Thailand*
Alistair Swale, *University of Waikato, New Zealand*
Naomi Yamashita, *NTT, Japan*
Yutaka Yamauchi, *Kyoto University, Japan*
Nadia Berthouze, *UCLIC, University College London, United Kingdom*

External Reviewers

Culture and Computing 2017

Atsushi Matsumura, *University of Tsukuba, Japan*
Kohei Furukawa, *Ritsumeikan University, Japan*
Liang Li, *Ritsumeikan University, Japan*
Norihiko Uda, *University of Tsukuba, Japan*
Takanobu Nishiura, *Ritsumeikan University, Japan*
Takeshi Miura, *Akita University, Japan*

Plan Optimization for Creating Bilingual Dictionaries of Low-Resource Languages

Arbi Haza Nasution
Department of Social Informatics,
Kyoto University
Kyoto, Japan
Email: arbi@ai.soc.i.kyoto-u.ac.jp

Yohei Murakami
Unit of Design,
Kyoto University
Kyoto, Japan
Email: yohei@i.kyoto-u.ac.jp

Toru Ishida
Department of Social Informatics,
Kyoto University
Kyoto, Japan
Email: ishida@i.kyoto-u.ac.jp

***Abstract*—The constraint-based approach has been proven useful for inducing bilingual lexicons for closely-related low-resource languages. When we want to create multiple bilingual dictionaries linking several languages, we need to consider manual creation by bilingual language experts 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. We adopt the Markov Decision Process (MDP) in formalizing plan optimization for creating bilingual dictionaries; the goal is to better predict the most feasible optimal plan with the least total cost before fully implementing the constraint-based bilingual dictionary induction framework. We define heuristics based on input language characteristics to devise a baseline plan for evaluating our MDP-based approach with total cost as an evaluation metric. The MDP-based proposal outperformed heuristic planning on total cost for all datasets examined.**

1. Introduction

Nowadays, machine-readable bilingual dictionaries are being utilized in actual services [1] [2] to support intercultural collaboration [3], but low-resource languages lack such sources. Previous work on high-resource languages showed the effectiveness of parallel corpora [4] [5] and comparable corpora [6] [7] in inducing bilingual lexicons. Clearly bilingual lexicon extraction is highly problematic for low-resource languages due to the paucity or outright omission of parallel and comparable corpora. We recently introduced the promising approach of treating pivot-based bilingual lexicon induction for low-resource languages as an optimization problem [8]. Bilingual dictionaries are the only language resource required by our approach. 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 lexicon induction approach, we need to consider the inclusion of more traditional methods like manually creating the bilingual dictionaries by bilingual language experts. Despite 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, plan optimization is required. We address the following research goals:

- *Formalization of plan optimization in creating bilingual dictionaries*: Modeling bilingual dictionary dependency with AND/OR graphs, and employing the Markov Decision Process for plan optimization.
- *Creating heuristic plans as baselines for evaluating plans*: We create plans that follow some basic heuristics and then use total cost as the evaluation metric for comparing performance.

The rest of this paper is organized as follows: In Section 2, we will briefly discuss related research on pivot-based bilingual dictionary induction. Section 3 discusses how to model dictionary dependency. Section 4 details the formalization of plan optimization, a core component of our proposal. Section 5 describes our experiments and the results. Finally, Section 6 concludes this paper.

2. Pivot-Based Bilingual Lexicon Induction

The first work on bilingual lexicon induction to create bilingual dictionaries (language A and language C) via pivot language B is Inverse Consultation (IC) [9]. It utilizes the structure of input dictionaries to measure the closeness of word meanings and then uses the results to prune erroneous translation pair candidates. The IC approach identifies equivalent candidates of language A words in language C by consulting dictionary A-B and dictionary B-C. These equivalent candidates will be looked up and compared in the inverse dictionary C-A.

The pivot-based approach is very suitable for low-resource languages, especially when dictionaries are the only language resource required. 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 pair candidates. To overcome this limitation, our team [10] proposed to treat pivot-based bilingual lexicon induction as an optimization

978-1-5386-1135-7/17 $31.00 © 2017 IEEE
DOI 10.1109/Culture.and.Computing.2017.21

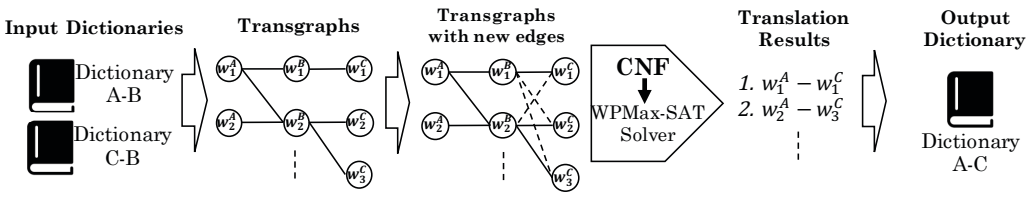


Figure 1. One-to-one constraint approach to pivot-based bilingual dictionary induction.

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). This assumption yielded the development of a constraint optimization model to induce an Uyghur-Kazakh bilingual dictionary using Chinese language as the pivot, which means that Chinese words were used as intermediates to connect Uyghur words in an Uyghur-Chinese dictionary with Kazakh words in a Kazakh-Chinese dictionary. The proposal uses a graph whose vertices represent words and edges indicate shared meanings; following [11] 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 A, w_1^B and w_2^B are pivot words in language B, and w_1^C, w_2^C and w_3^C are non-pivot words in language 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 [12] 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 translation pairs needed to offset resource paucity because few such pairs can be found. Therefore, we generalized the constraint-based bilingual lexicon 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 [8].

3. Modeling Dictionary Dependency

Our constraint-based bilingual dictionary induction requires two bilingual dictionaries that share the same pivot language. We can induce bilingual dictionary A-C from bilingual dictionary A-B and B-C as input (language B is the pivot). Nevertheless, we can also induce bilingual dictionary A-C with different input bilingual dictionaries using language D as the pivot language for instance. We use an AND/OR graph to model the dependency: bilingual dictionary A-C can be induced from bilingual dictionaries A-B AND B-C OR from bilingual dictionaries A-D AND D-C as shown in Figure 2.

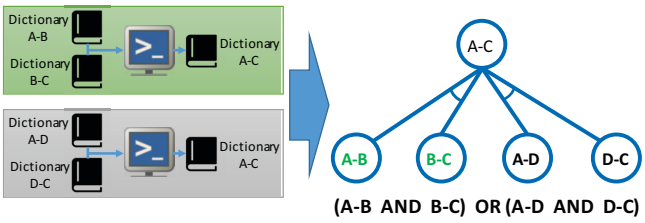


Figure 2. Modeling Bilingual Dictionary Induction Dependency.

If two sets of input dictionaries can be used to induce bilingual dictionary A-C, if we have to choose between the two sets, we need to prioritize input dictionaries that can induced bilingual dictionary A-C with more correct translation pairs. But, the number of correct translation pairs that can be induced depends on the quality of the constraint-based bilingual dictionary induction which also depends on the size and quality of the input dictionaries which can best be treated as a search problem with uncertainty.

4. Formalizing Plan Optimization

We assume that user provides the number of existing translation pairs for existing bilingual dictionaries. We calculate language similarity between each language pair with ASJP. Both characteristics are crucial in predicting the performance of constraint-based bilingual dictionary induction which also determines the number of translation pairs in the output bilingual dictionary. User can also set the minimum number of translation pairs the output bilingual dictionary should have. Markov Decision Process (MDP) is a well-known technique to solve problems containing uncertainty. An MDP is the tuple (S, A, $T_{s,a,s'}$, $C_{s,a,s'}$), where S is a set of states, A is a set of actions, $T_{s,a,s'}$ is a transition probability distribution over the state space when action a is taken in state s, and $C_{s,a,s'}$ is the negative reward or cost for taking action a in state s.

4.1. State

If n is a number of languages of interest specified by user, the total number of all possible combinations of bilingual dictionaries in the state is $m = C_2^n$. For example, if we have 4 languages (L_1, L_2, L_3, L_4), there will be $m = C_2^4 = 6$ bilingual dictionaries.

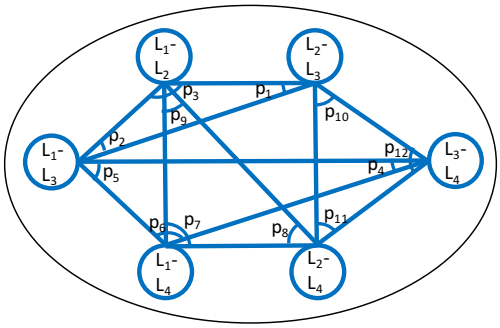


Figure 3. Bilingual Dictionary Induction Dependency Model.

Each state stores m bilingual dictionaries, each $D_{(x,y)}$ with four possible status types: not existing $D_{(x,y):n}$, existing but number of translation pairs is below user requested size $D_{(x,y):eu}$, induced from pivot action with z as pivot language but the number of translation pairs is below user request $D_{(x,y):pu(z)}$, and existing or induced from pivot action where the number of translation pairs equals or exceeds user requested number of translation pairs $D_{(x,y):s}$, hence, the maximum number of MDP states is $4^m = 4^6 = 4,096$. Based on the status, we further categorize the bilingual dictionary as either SATDict ($D_{(x,y):s}$) or UnSATDict (($D_{(x,y):n}$, $D_{(x,y):eu}$, or $D_{(x,y):pu(z)}$)). Each state also stores information about the dependency between its dictionaries as shown in Figure 3.

After an agent takes an action in state s to enrich an UnSATDict of language x and y, if the size of the output dictionary satisfies user request, the agent will transit to the next one step ahead state, s', which has an SATDict of the same languages, x and y, while the other bilingual dictionaries in s' are unchanged from the previous state, s. State s' takes its state, SATState or UnSATState, from that of state s, as shown in Figure 4.

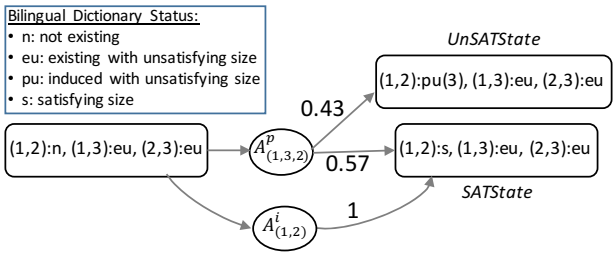


Figure 4. Example of State Transition to SATState or UnSATState.

The number of states increases exponentially with the number of languages as shown in Table 1. So as to cast formulation complexity into a graph theory problem, we initially create only one Start-State where each bilingual dictionary status is calculated from the input bilingual dictionaries size given by user. We then generate all possible actions, including the next one step ahead states s'_k that would result from taking those actions, based on the bilingual dictionary induction dependency model as shown in Figure 3. We iterate the same procedure on s'_k until the Final-State is reached where all m bilingual dictionaries from n languages are available where the number of translation pairs equals or exceeds user requested number of translation pairs.

TABLE 1. EXPONENTIAL RELATIONSHIP BETWEEN NUMBER OF LANGUAGES AND STATES

#Languages (n)	#Dictionaries (m)	#states
2	1	4
3	3	64
4	6	4,096
5	10	1,048,576
6	15	1,073,741,824

4.2. Action

Some bilingual dictionary creation methods require only bilingual dictionaries as input, and thus they can be used in this MDP model such as the Inverse Consultation method, the one-to-one constraint-based approach, and our generalized constraint-based bilingual lexicon induction framework. However, since our method outperformed both previous methods, we apply our method as one of MDP action and call it pivot action $A^p_{(x,z,y)}$ where z is the pivot language. As we assume low-resource language, adequate machine-readable bilingual dictionaries will be unavailable, and we define manual bilingual dictionary creation by a language expert as investment action $A^i_{(x,y)}$. The purposes of the pivot action and investment action are to enrich and change the category of the bilingual dictionaries stored in each state from UnSATDict to SATDict. An UnSATDict with status $D_{(x,y):n}$ or $D_{(x,y):eu}$ can be enriched by both investment action and pivot action. For an UnSATDict with status $D_{(x,y):pu(z)}$, however, we limit the next action to investment action only because pivot action $A^p_{(x,z,y)}$ was already tried exactly one step prior. A pivot action can be taken from input dictionaries with status $D_{(x,y):s}$ and $D_{(x,y):eu}$.

4.3. Cost

In the MDP model, the agent expects to get a reward after taking some actions. The reward will guide the agent to reach the final state and obtain the best path or in this case the best plan. Because for creating a bilingual dictionary we need to pay some cost instead of getting some rewards afterward, here we cast the reward as a cost. The terms of reward and cost are interchangeable in many previous MDP studies [13]. When we take an investment action, we are actually asking a language expert to manually create a bilingual dictionary and we need to pay for the time and effort incurred. For the sake of modeling simplicity, we take \$0.3 to be the cost of the investment action for one translation pair. For taking pivot action, i.e., using the constraint-based bilingual lexicon induction framework, when we already

have the input dictionaries, we can generate the output dictionary in a short time. Thus, we assume that there is no cost in taking the pivot action.

4.4. State Transition Probability

The number of translation pairs in the current state affects the performance of the pivot action taken in the current state, and thus the number of induced translation pairs in the next state. When the bilingual dictionary output by the pivot action $A^p_{(x,z,y)}$ equals or exceeds user requested size, the agent will transit to the next state, s' in which the bilingual dictionary status of languages x and y is $D_{(x,y):s}$ or else transit to the next state, s' in which the bilingual dictionary status of languages x and y is $D_{(x,y):pu(z)}$ and the remaining bilingual dictionaries in the state s' are unchanged from the previous state s.

The state transition probability for taking a pivot action depends on the size of output bilingual dictionary which also depends on the precision of the constraint-based bilingual lexicon induction framework. If the precision is 1, then all translation pair candidates are taken as translation pairs. The experiment results in our previous work [8] showed a trend that precision is greater than input language similarity and we can obtain at least as many translation pair candidates as number of entries of the smaller of the input bilingual dictionaries. For the sake of simplicity, we assume that we can obtain as many translation pair candidates as the number of entries of the smaller input bilingual dictionaries and that the precision of constraint-based bilingual lexicon induction equals or exceeds input languages similarity. Nevertheless, since we do not know the exact precision for the different input language sets before implementing the constraint-based bilingual lexicon induction and evaluating the result, we do not know the exact number of translation pairs of the evaluated output bilingual dictionary $EPair$. Therefore, we calculate the probability of the transition to the SATState by Equation 1 where SATRange is the range of precision that will ensure that the output dictionary size will satisfy user request while UnSATRange is the range of precision that will not satisfy user requested size. We use Equations 2 and 3 to calculate $SATRange$ and $UnSATRange$, respectively, given precision $SATPrec$, i.e., the minimum precision required to satisfy user request following Equation 4. We assume that output bilingual dictionary size is linear to pivot action precision as shown in Figure 5.

$$T_{s,a,s'} = SATRange/(SATRange + UnSATRange) \quad (1)$$

$$SATRange = maxPrec - SATPrec \quad (2)$$

$$UnSATRange = SATPrec - minPrec \quad (3)$$

$$SATPrec = RPair/CPair \quad (4)$$

For instance, when we want to enrich UnSATDict $D_{(1,2):n}$ with pivot action $A^p_{(1,3,2)}$ from existing UnSATDict $D_{(1,3):eu}$ with input dictionary size equals $6,000$ and $D_{(2,3):eu}$ with input dictionary size equals $8,500$, we can

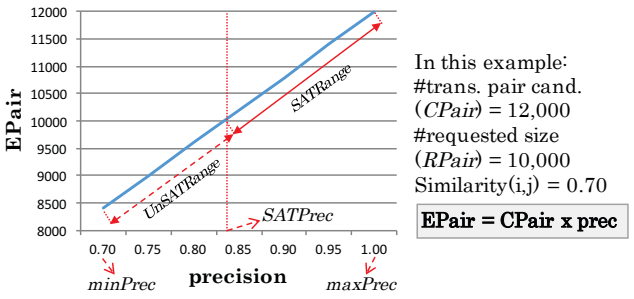


Figure 5. Example of Linear Relationship between Output Bilingual Dictionary's Size and Precision.

get the number of translation pair candidates ($CPair$) that equals the minimum input dictionary size which is $6,000$. If we also get information about user requested size of output bilingual dictionary ($RPair$) which is $5,000$ with 0.70 language similarity, we can assume the minimum precision ($minPrec$) of the pivot action is also 0.70 and the maximum precision ($maxPrec$) is 1. In this example, $SATPrec = 5,000/6,000 = 0.83$, and we can further calculate $SATRange = 1 - 0.83 = 0.17$ and $UnSATRange = 0.83 - 0.70 = 0.13$. Thus, the transition probability when taking action a which in this case $A^p_{(1,3,2)}$ in state s which has UnSATDict $D_{(1,2):n}$ to a SATState s'_{SAT} which has SATDict $D_{(1,2):s}$ is $T_{s,a,s':SAT} = 0.17/(0.17 + 0.13) = 0.57$. Consequently, the transition probability to UnSATState s'_{UnSAT} which has UnSATDict $D_{(1,2):pu(3)}$ is $T_{s,a,s':UnSAT} = 1 - T_{s,a,s':SAT} = 1 - 0.57 = 0.43$. To enrich the bilingual dictionary $D_{(1,2):n}$ with investment action $A^i_{(1,2)}$, we assume that the language expert can always satisfy user requested output bilingual dictionary size, thus the agent will transit to SATState with probability of 1. This state transition example is depicted in Figure 4.

4.5. Value Iteration

We use value iteration algorithm [14] to calculate utility (optimal policy) of each state by summing the cost for starting at state s and acting according to policies thereafter as shown in Equation 5. Every state will have a policy of best action in order to minimize cumulative costs.

$$U_{i+1}(s) = min_{a \in A(s)} \sum_{s'} T_{s,a,s'} \left(C_{s,a,s'} + U_i(s') \right) \quad (5)$$

5. Experiment

Automated Similarity Judgment Program (ASJP) was proposed by [15] with the main goal of developing a database of Swadesh lists [16] for all of the world's languages from which lexical similarity or lexical distance matrix between languages can be obtained by comparing the word lists. We utilize ASJP to select the target languages used in our simulation case studies. Indonesia has 707 low-resource ethnic languages [17], all of which are suitable

TABLE 2. SIMILARITY MATRIX OF 6 INDONESIAN ETHNIC LANGUAGES RANKED BY NUMBER OF SPEAKERS

Language	Indonesian	Old Javanese	Minangkabau	Makasar	Sasak
Old Javanese	0.2409				
Minangkabau	0.6159	0.2501			
Makasar	0.3305	0.2239	0.3307		
Sasak	0.4391	0.2045	0.4439	0.3507	
Lampung	0.2070	0.1960	0.1910	0.1990	0.2079

TABLE 3. EXISTING PRINTED BILINGUAL DICTIONARY SIZE

Bilingual Dictionary	Number of Translation Pair
Indonesian - Old Javanese	8,600
Indonesian - Minangkabau	12,600
Indonesian - Makasar	4,300
Indonesian - Sasak	19,300
Indonesian - Lampung	9,200

as target languages in our study. From them we selected Indonesian, Old Java, Minangkabau, Makasar, Sasak, and Lampung. We then generated the language similarity matrix by utilizing ASJP as shown in Table 2. There is no available dictionary either machine readable or printed format between these Indonesian ethnic languages. Nevertheless, we can find several printed bilingual dictionaries for Indonesian (the official language) and several ethnic languages as shown in Table 3. Actually we need to convert those printed dictionaries to machine readable dictionaries either manually by human or automatically as shown in [18]. For test simplicity, we assumed that all printed dictionaries had been converted into machine readable form.

5.1. Experiment Settings

To show effectiveness of our method, we use, as baselines, heuristic plans to choose an action to take based on input language characteristics as follows:

A. Heuristic 1:

- Only investment actions are considered.

B. Heuristic 2:

- Prioritizing pivot actions over investment actions as the first priority.
- Prioritizing input bilingual dictionary size (bigger is better) as the second priority.
- Prioritizing language similarity (closer is better) as the third priority.
- Before taking the selected pivot action, an investment action can be taken to ensure that the input dictionaries specified by the selected pivot action satisfying user requested size.

C. Heuristic 3:

- Prioritizing pivot actions over investment actions as the first priority.
- Prioritizing language similarity (closer is better) as the second priority.
- Prioritizing input bilingual dictionary size (bigger is better) as the third priority.
- Before taking the selected pivot action, an investment action can be taken to ensure that the input dictionaries specified by the selected pivot action satisfying user requested size.

We conducted experiments on four datasets from the printed dictionaries of Indonesian ethnic languages. We randomly selected language pairs to get two four-languages datasets and two five-languages datasets. Dataset 1 had four languages: (1) Indonesia, (2) Minangkabau, (3) Lampung, and (4) Sasak. Dataset 2 also had four languages: (1) Indonesia, (2) Minangkabau, (3) Makasar, and (4) Old Java. Dataset 3 had five languages: (1) Indonesia, (2) Minangkabau, (3) Lampung, (4) Sasak, and (5) Makasar. Dataset 4 also had five languages: (1) Indonesia, (2) Minangkabau, (3) Lampung, (4) Makasar, and (5) Old Java. Plans were evaluated by comparing the total cost of MDP and heuristic-based optimal plan.

5.2. Experiment Result

The heuristic approach that uses only the language characteristics with rules to follow fails to find the optimal path with the least cost. The language characteristics should be processed into more useful information like both cost of action and transition probability as in the proposed MDP approach to get the feasible optimal plan with the least cost. The result depicted in Table 4 shows that our MDP model outperformed the heuristic plans as regards total cost of the optimal plan for all datasets. Compared to heuristic 1, heuristic 2 and heuristic 3 plans, respectively, our MDP plan reduced total cost by, for dataset 1, 75%, 65%, and 68%; for dataset 2, 59%, 47%, and 47%; for dataset 3, 77%, 68%, and 68%, and for dataset 4, 62%, 51%, and 51%.

Our recursive algorithm greatly reduced the complexity of the MDP model as shown in Table 5. In datasets 1 and 2, both of which held 4 languages, the number of states was reduced by 93% compared to the maximum number of states shown in Table 1. For datasets 3 and 4, which had 5 languages, the reduction was 99%.

6. Conclusion

Our constraint-based bilingual lexicon induction approach has the potential to enrich low-resource languages

TABLE 4. TOTAL COST COMPARISON BETWEEN MDP AND HEURISTICS APPROACH

Dataset	Method	Planning	Total Cost
1	Heuristic 1	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(3,4)}$	$23,670
	Heuristic 2	$A^i_{(1,4)} \rightarrow A^i_{(1,2)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^p_{(3,1,4)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(3,4)}$	$16,937
	Heuristic 3	$A^i_{(1,4)} \rightarrow A^i_{(1,2)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^p_{(3,2,4)} \rightarrow A^i_{(1,3)} \rightarrow A^i_{(3,4)}$	$18,515
	MDP	$A^i_{(1,3)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(2,4)} \rightarrow A^p_{(1,4,2)} \rightarrow A^p_{(3,1,4)} \rightarrow A^p_{(1,2,4)} \rightarrow A^p_{(2,4,3)}$	$5,884
2	Heuristic 1	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(3,4)}$	$28,350
	Heuristic 2	$A^i_{(1,2)} \rightarrow A^i_{(1,4)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^p_{(3,1,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(3,4)}$	$21,992
	Heuristic 3	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^p_{(3,2,4)} \rightarrow A^i_{(3,4)}$	$21,973
	MDP	$A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^p_{(3,1,4)} \rightarrow A^p_{(2,1,3)} \rightarrow A^i_{(2,3)} \rightarrow A^p_{(1,3,2)} \rightarrow A^p_{(2,3,4)}$	$11,629
3	Heuristic 1	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^i_{(3,4)} \rightarrow A^i_{(3,5)} \rightarrow A^i_{(4,5)}$	$46,380
	Heuristic 2	$A^i_{(1,4)} \rightarrow A^i_{(1,2)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^p_{(3,1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(2,1,5)} \rightarrow A^p_{(4,1,5)} \rightarrow A^p_{(3,1,5)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^i_{(4,5)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(3,4)} \rightarrow A^i_{(3,5)}$	$33,380
	Heuristic 3	$A^i_{(1,4)} \rightarrow A^i_{(1,2)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(2,1,5)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^p_{(4,2,5)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^p_{(3,1,4)} \rightarrow A^i_{(4,5)} \rightarrow A^i_{(3,4)} \rightarrow A^p_{(3,4,5)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(3,5)}$	$33,325
	MDP	$A^i_{(1,3)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(2,4)} \rightarrow A^p_{(1,4,2)} \rightarrow A^p_{(3,1,4)} \rightarrow A^p_{(2,4,3)} \rightarrow A^p_{(4,1,5)} \rightarrow A^p_{(1,2,4)} \rightarrow A^p_{(2,4,5)} \rightarrow A^p_{(3,2,5)}$	$10,594
4	Heuristic 1	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^i_{(3,4)} \rightarrow A^i_{(3,5)} \rightarrow A^i_{(4,5)}$	$49,590
	Heuristic 2	$A^i_{(1,2)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(2,1,5)} \rightarrow A^p_{(3,1,5)} \rightarrow A^i_{(1,4)} \rightarrow A^p_{(2,1,4)} \rightarrow A^p_{(4,1,5)} \rightarrow A^p_{(3,1,4)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^i_{(2,3)} \rightarrow A^i_{(4,5)} \rightarrow A^i_{(3,4)} \rightarrow A^i_{(3,5)}$	$38,443
	Heuristic 3	$A^i_{(1,2)} \rightarrow A^i_{(1,4)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(2,1,5)} \rightarrow A^i_{(1,3)} \rightarrow A^p_{(2,1,3)} \rightarrow A^i_{(2,4)} \rightarrow A^i_{(2,5)} \rightarrow A^p_{(4,2,5)} \rightarrow A^p_{(3,1,4)} \rightarrow A^p_{(3,1,5)} \rightarrow A^i_{(4,5)} \rightarrow A^i_{(3,4)} \rightarrow A^i_{(3,5)} \rightarrow A^i_{(2,3)}$	$38,424
	MDP	$A^i_{(1,3)} \rightarrow A^i_{(1,4)} \rightarrow A^i_{(1,5)} \rightarrow A^p_{(3,1,4)} \rightarrow A^p_{(4,1,5)} \rightarrow A^p_{(4,5)} \rightarrow A^p_{(3,4,5)} \rightarrow A^p_{(2,1,4)} \rightarrow A^i_{(2,4)} \rightarrow A^p_{(1,4,2)} \rightarrow A^p_{(2,4,3)} \rightarrow A^p_{(2,3,5)}$	$18,761

TABLE 5. MDP MODEL COMPLEXITY REDUCTIONS

Dataset	#Languages	#States	#Invest. Action	#Pivot Action
1 & 2	4	280	408	432
3 & 4	5	8,704	12,800	15,360

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. 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. 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 key research contribution is a formalization of planning optimization for creating bilingual dictionaries. 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. To the best of our knowledge, our study is the first attempt to formalize planning optimization for creating bilingual dictionaries.

Acknowledgments

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 Endownment Fund for Education (LPDP).

References

[1] T. Ishida, "Language grid: An infrastructure for intercultural collaboration," in *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, 2006, pp. 96–100.

[2] T. Ishida, Ed., *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer Publishing Company, Incorporated, 2011.

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

[4] P. Fung, "A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora," in *Machine Translation and the Information Soup*. Springer, 1998, pp. 1–17.

[5] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[6] R. Rapp, "Identifying word translations in non-parallel texts," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 320–322.

[7] P. Fung, "Compiling bilingual lexicon entries from a non-parallel english-chinese corpus," in *Proceedings of the Third Workshop on Very Large Corpora*, 1995, pp. 173–183.

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

[9] K. Tanaka and K. Umemura, "Construction of a bilingual dictionary intermediated by a third language," in *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1994, pp. 297–303.

[10] M. Wushouer, D. Lin, T. Ishida, and K. Hirayama, "A constraint approach to pivot-based bilingual dictionary induction," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 15, no. 1, pp. 4:1–4:26, Nov. 2015. [Online]. Available: http://doi.acm.org/10.1145/2723144

[11] S. Soderland, O. Etzioni, D. S. Weld, M. Skinner, J. Bilmes *et al.*, "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*. Association for Computational Linguistics, 2009, pp. 262–270.

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

[13] D. J. White, "A survey of applications of markov decision processes," *Journal of the Operational Research Society*, vol. 44, no. 11, pp. 1073–1096, 1993.

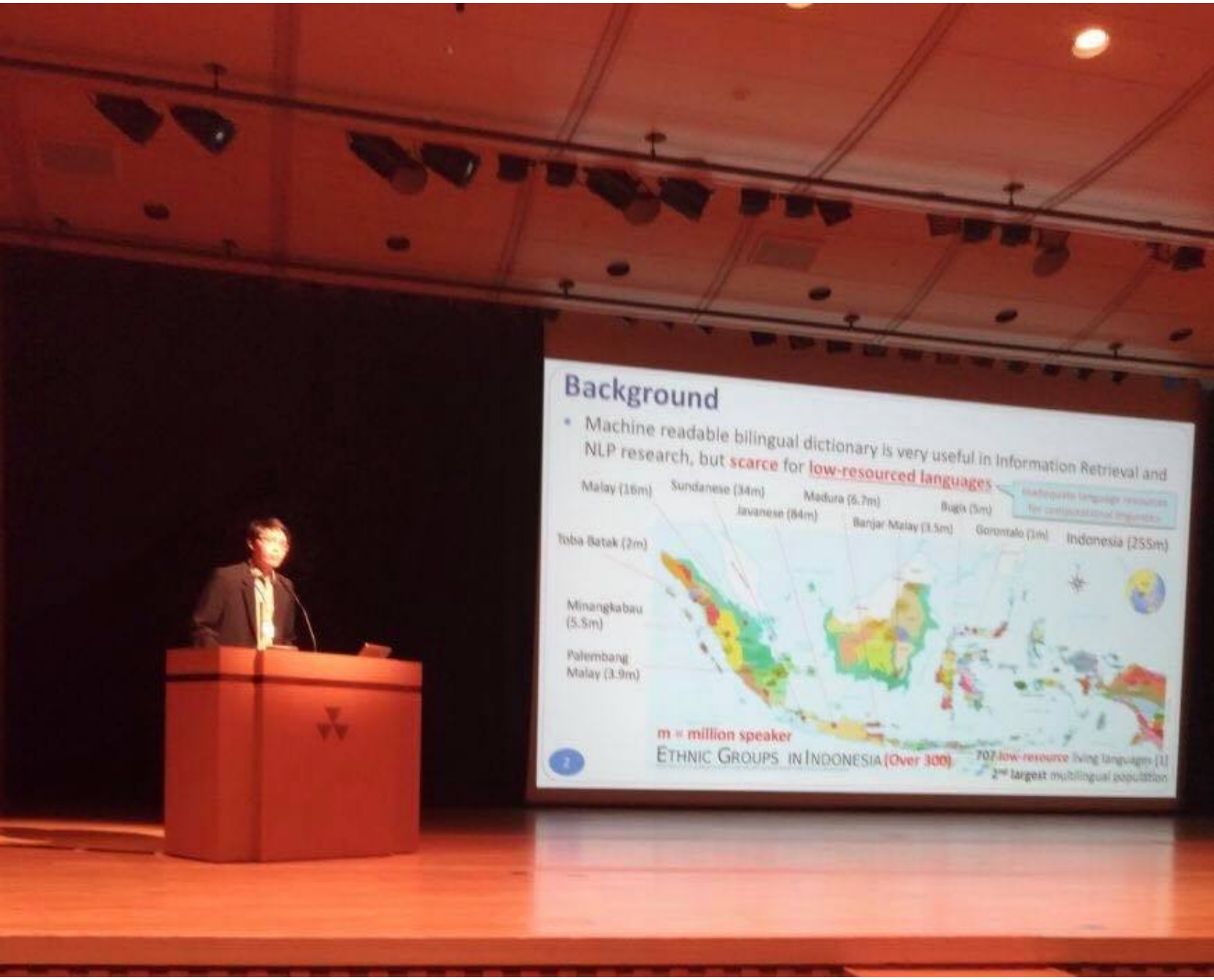
[14] R. A. Howard, *Dynamic Programming and Markov Processes.* The M.I.T. Press, 1960.

[15] E. W. Holman, C. H. Brown, S. Wichmann, A. Müller, V. Velupillai, H. Hammarström, S. Sauppe, H. Jung, D. Bakker, P. Brown *et al.*, "Automated dating of the world's language families based on lexical similarity," *Current Anthropology*, vol. 52, no. 6, pp. 841–875, 2011.

[16] M. Swadesh, "Towards greater accuracy in lexicostatistic dating," *International journal of American linguistics*, vol. 21, no. 2, pp. 121–137, 1955.

[17] M. P. Lewis, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 18th ed. Dallas, Texas: SIL International, 2015. [Online]. Available: http://www.ethnologue.com

[18] J. Mayfield and P. McNamee, "Converting on-line bilingual dictionaries from human-readable to machine-readable form," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 405–406.


Background
Machine readable bilingual dictionary is very useful in Information Retrieval and NLP research, but scarce for low-resourced languages
Malay (16m)
Sundanese (34m)
Madura (6.7m)
Javanese (84m)
Bugis (5m)
Banjar Malay (3.5m)
Gorontalo (1m)
Indonesia (255m)
Toba Batak (2m)
Minangkabau (5.5m)
Palembang Malay (3.9m)
m = million speaker
ETHNIC GROUPS IN INDONESIA (Over 300)
707 low-resource living languages (1)
2nd largest multilingual population
2