Pivot-based hybrid machine translation to support multilingual communication for closely related languages

Arbi H. Nasution

Universitas Islam Riau
Pekanbaru, Indonesia

ABSTRACT: Machine translation (MT) is very useful in supporting multicultural communication. Existing statistical machine translation (SMT), which requires high quality and quantity of corpora, and rule-based machine translation (RBMT), which requires bilingual dictionaries, morphological, syntax and semantic analysers, are scarce for low-resource languages. Due to the lack of language resources, it is difficult to create MT from high-resource languages to low-resource languages, such as Indonesian ethnic languages. Nevertheless, due to Indonesian ethnic languages' characteristics, a pivot-based hybrid machine translation (PHMT) can be introduced by combining SMT and RBMT with Indonesian as a pivot, which then can be utilised in a multilingual communication support system. The PHMT translation quality was evaluated, with fluency and adequacy as metrics, and then the usability of the system was evaluated. Despite the medium average translation quality (3.05 fluency score and 3.06 adequacy score), the 3.71 average mean score of the usability evaluation indicates that the system is usable to support multilingual collaboration.

INTRODUCTION

Machine translation (MT) is very useful in supporting multicultural communication, but scarce for low-resource languages. Existing MT research uses statistical machine translation (SMT), which requires high quality and quantity of corpora, and rule-based machine translation (RBMT), which requires bilingual dictionaries, a morphological analyser, syntax analyser (parser) and semantic analyser.

There are research challenges in creating MT from high-resource languages (HRL) to low-resource languages (LRL), such as Indonesian ethnic languages. These lack an adequate corpora, sizable dictionary, good morphological, syntax and semantic analysers. Nevertheless, Indonesian ethnic languages characteristics with several clusters of similar languages having similar morphology and syntax provide a good starting point to address these challenges. The following research goals were addressed:

1. To develop pivot-based hybrid machine translation (PHMT). This is a combination of SMT and RBMT and aims to bridge the gap between HRLs and LRLs.

2. Support multilingual communication with the PHMT by the implementation of the PHMT to develop a multilingual communication support system.

CLOSELY RELATED LANGUAGES

Historical linguistics is the scientific study of language change over time in term of sound, analogical, lexical, morphological, syntactic and semantic information [1]. Comparative linguistics is a branch of historical linguistics that is concerned with language comparison to determine historical relatedness and to construct language families [2].

Many methods, techniques and procedures have been utilised in investigating the potential distant genetic relationship of languages, including lexical comparison, sound correspondences, grammatical evidence, borrowing, semantic constraints, chance similarities and sound-meaning isomorphism [3]. The genetic relationship of languages is used to classify languages into language families. Closely related languages are those that came from the same origin or proto-language and belong to the same language family.

Glottochronology is one lexical comparison method for estimating the amount of time elapsed since related languages diverged from a common ancestral language [4]. Glottochronology depends on a basic, relatively culture-free vocabulary, which is known as a Swadesh list. The automated similarity judgment program (ASJP) [5] has the main goal of developing a database of Swadesh lists for all of the world's languages from which lexical similarity or a lexical

distance matrix between languages can be obtained by comparing the word lists [4]. For example, Indonesia has 707 low-resource ethnic languages, which mostly belong to the same language family, i.e. the Austronesian language family [6]. The language similarity matrix can be generated by utilising the ASJP.

Closely related languages share cognates with common semantics or meaning of the lexicons [2]. Some linguistic studies show that the percentage of shared cognates, either related directly or via a synonym, constitutes a highly accurate linguistic distance measure based on mutual intelligibility, i.e. the ability of speakers of one language to understand the other language [7][8]. The higher the percentage of shared cognates between the languages, the lower the linguistic distance and the higher is the level of mutual intelligibility.

BILINGUAL LEXICON INDUCTION

Machine readable bilingual lexicons are very useful for natural language processing applications/research, such as cross-language information retrieval [9] and machine translation [10], but are usually unavailable for low-resource languages. These lexicons traditionally are extracted from parallel corpora, a corpus that contains source texts and their translations. Various techniques are used to extract bilingual lexicons from parallel corpora other than the traditional sentence-aligned bilingual texts [10].

A better method of producing word alignment is by training inversion transduction grammars [11], while recently English monolingual semantic role labelling was utilised to obtain more semantically correct bilingual correlations [12]. An inductive chain learning method can even automatically acquire bilingual rules from parallel corpora without utilising bilingual dictionary or machine translation [13].

However, despite good results in the extraction of bilingual lexicons, parallel corpora remain scarce resources for low-resource languages. Thus, research in bilingual lexicon extraction has shifted to comparable corpora consisting of texts sharing common features, such as domain, genre, register or sampling period without having a source text-target text relationship [14-16]. The approach depends on the assumption that the term and its translation appear in similar contexts, which means that a translation equivalent of a source word can be found by identifying a target word with the most similar context vector in a comparable corpus [14][15].

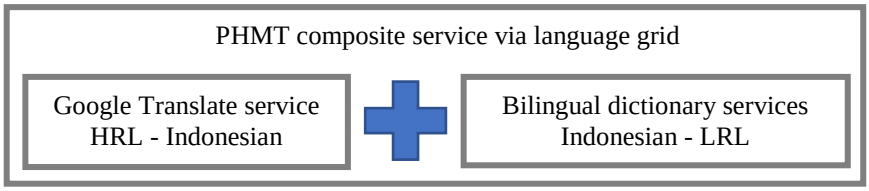
Identification of good similarity metrics as signals of translation equivalence is the main research challenge in this area. A discriminative model of bilingual lexicon induction from comparable corpora was presented and showed good experimental results on a wide variety of languages (many of them low-resource), for which a wide variety of monolingual corpora and seed bilingual dictionaries are available [17].

Nevertheless, bilingual lexicon extraction is still highly problematic for most low-resource languages, due to the paucity or outright omission of parallel and comparable corpora. Recent research on creating bilingual dictionaries of Indonesian ethnic languages collaboratively with native speakers [18] shows a great potency by following the plan optimiser [19] and utilising a constraint-based bilingual lexicon induction by creating bilingual dictionary A-C with only bilingual dictionaries A-B and B-C as input [20][21]. The output machine readable bilingual dictionary was wrapped as a service in Language Grid [22] to support intercultural collaboration [23].

PIVOT-BASED HYBRID MACHINE TRANSLATION

Extending previous work [24], Google Translate service and bilingual dictionary service were combined as a composite service in the language grid, as shown in Figure 1. There are more than a hundred high-resource languages available in the Google Translate service. To this date, two Indonesian ethnic languages, i.e., Javanese and Sundanese, are available in Google Translate service alongside the official language, Indonesian.

It is unlikely that Google Translate can provide the rest of Indonesian ethnic languages in the near future, since the available corpora for Indonesian ethnic languages are still scarce. In order to bridge the gap between high-resource languages and low-resource languages, in this case between English and Minangkabau, a quicker approach is to create an English-Minangkabau PHMT with Indonesian as the pivot (see Figure 1). Since Minangkabau has 61.59% lexical similarity with Indonesian based on ASJP, the morphology and syntax are similar. Therefore, Indonesian-Minangkabau word-to-word translation is expected to be acceptable.



Note: HRL: high-resource languages; LRL: low-resource languages

Figure 1: PHMT as a language grid composite service.

EXPERIMENT

A multi-language support system for international symposia has been provided by combining human inputters and language services [25]. The PHMT was used to support multilingual communication. In this experiment, the system supported a Minangkabau-speaking audience in understanding an English presentation. The Indonesian-Minangkabau dictionary service used in this research has 5,391 entries.

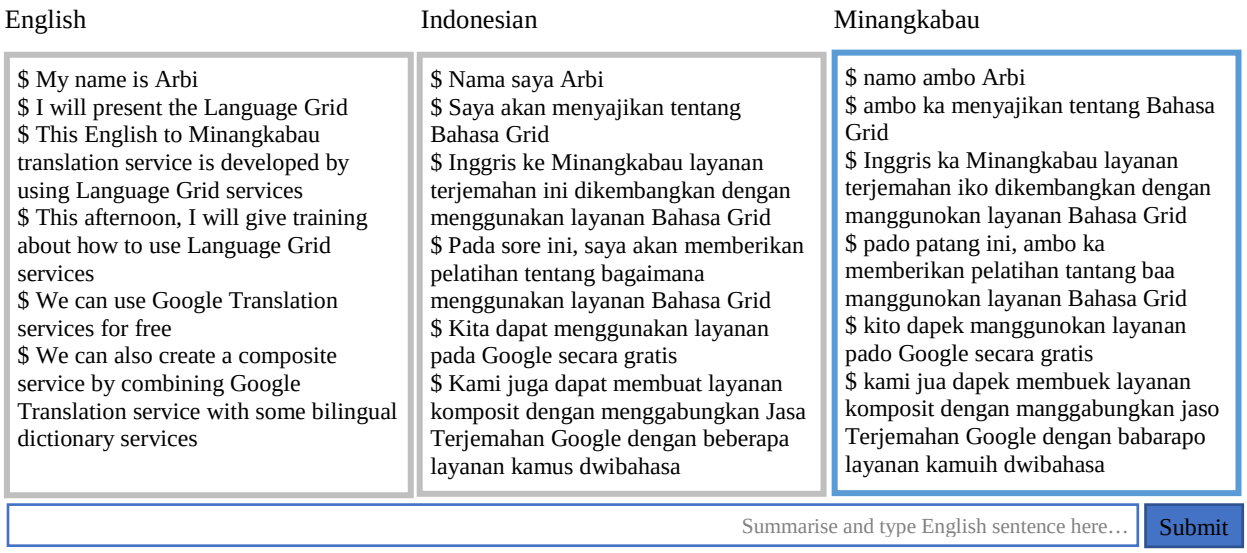


Figure 2: PHMT input screen.

Figure 3: PHMT client screen.

A video of an English presentation was played to 165 Bachelor of informatics students of the Islamic University of Riau, Indonesia. The video and the system were displayed on a separate screen. While listening to the English presentation, a simplified English sentence was input to the system, as shown in Figure 2. Audiences could view the system from any Web browser (personal PC or smartphone), as shown in Figure 3.

RESULTS

The translation quality of both English-Indonesian translations and Indonesian-Minangkabau translations were assessed and the usability of the multilingual communication support system evaluated.

Translation Quality Assessment

The translation quality was assessed with fluency and adequacy as measures following the linguistic data annotation specification [26] with a 5-point scale (1 for the lowest score to 5 for the highest). Fluency refers to the degree to which the translation is well formed according to the rules/grammar of the language. A fluent translation is one that is well-

formed grammatically, has correct spelling, using common terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of the language. The fluency of the English-Indonesian translations and Indonesian-Minangkabau translations were evaluated by bilingual speakers of those languages as judges, as shown in Table 1. The average fluency score of English-Indonesian translations and Indonesian-Minangkabau translations were 3.52 and 3.05, respectively.

Table 1: Fluency assessment.

How do you judge the fluency of this translation? It is:
5 - Flawless
4 - Good
3 - Non-native
2 - Disfluent
1 - Incomprehensible

Adequacy refers to the degree to which information in the original text is also conveyed in the translation. The adequacy of the English-Indonesian translations and Indonesian-Minangkabau translations were also evaluated by bilingual speakers of those languages as judges, as shown in Table 2. The judges determined whether the translation was adequate by comparing the English-Indonesian translations and Indonesian-Minangkabau translations against the reference translations. The average adequacy score of English-Indonesian translations and Indonesian-Minangkabau translations were 3.59 and 3.06, respectively.

Table 2: Adequacy assessment.

How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?
5 - All
4 - Most
3 - Much
2 - Little
1 - None

Usability Evaluation

The usability of the multilingual support system with the pivot-based hybrid machine translation was evaluated with a quantitative study using a questionnaire that consisted of seven items scaled from 1 (extreme - disagree) to 5 (extreme - agree). The average mean score was 3.71, as shown in Table 3. This result shows that the multilingual support system with the pivot-based hybrid machine translation is usable to support multilingual collaboration.

Table 3: Usability evaluation of PHMT.

Question	Proportion of each scale*					Mean
	1	2	3	4	5	
1. Was the interface easy to look at?	0.006	0.073	0.341	0.246	0.335	3.81
2. Did you understand the content of the presentation?	0.000	0.056	0.335	0.385	0.223	3.74
3. Was the Minangkabau translation result correct?	0.000	0.207	0.458	0.257	0.078	3.25
4. Was the Minangkabau translation result easy to understand?	0.000	0.089	0.464	0.307	0.140	3.51
5. Was the Minangkabau translation result helpful to understand the presentation?	0.006	0.078	0.458	0.291	0.168	3.56
6. Was the translation displayed in a timely manner?	0.006	0.045	0.296	0.346	0.307	3.96
7. Do you think this system is needed and important to support multilingual communication in international seminars between English native speakers with non-native audience?	0.006	0.022	0.251	0.257	0.464	4.11

* Scaled from 1 (extreme - disagree) to 5 (extreme - agree)

CONCLUSIONS

There were only small decreases of translation quality of the Indonesian-Minangkabau translations from the English-Indonesian translations of 13% for the average fluency score and 15% for the average adequacy score. Even though the fluency and adequacy scores are considered medium, the result is promising since only the simplest RBMT method was used, i.e. word-to-word translation of English-Indonesian translations to Minangkabau.

Based on the audience comments from the questionnaire, future work could improve the PHMT quality by refining and adding more entries to the Indonesian-Minangkabau Bilingual Dictionary, after consulting language experts. Multiple inputters could be used to improve translation speed and quality in supporting multilingual communication.

ACKNOWLEDGMENTS

This research was partially supported by Universitas Islam Riau. The author was supported by the Indonesia Endowment Fund for Education (LPDP).

REFERENCES

1. Campbell, L., *Historical Linguistics*. Edinburgh University Press (2013).
2. Lehmann, W.P., *Historical Linguistics: an Introduction*. Routledge (2013).
3. Campbell, L. and Poser, W.J., *Language Classification. History and Method*. Cambridge (2008).
4. Swadesh, M., Towards greater accuracy in lexicostatistic dating. *Inter. J. of American Linguistics*, 21, **2**, 121-137 (1955).
5. Holman, E.W., Brown, C.H., Wichmann, S., Muller, A., Velupillai, V., Hammarstrom, H., Sauppe, S., Jung, H., Bakker, D., Brown, P. and Belyaev, O., Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52, **6**, 841-875 (2011).
6. Lewis, M.P., Simons, G.F. and Fennig, C.D. (Eds), *Ethnologue: Languages of the World.* (18th Edn), SIL International, Dallas, Texas (2015).
7. Van Bezooijen, R. and Gooskens, C., How easy is it for speakers of Dutch to understand Frisian and Afrikaans, and why? *Linguistics in the Netherlands*, 22, **1**, 13-24 (2005).
8. Gooskens, C., Linguistic and extra-linguistic predictors of inter-Scandinavian intelligibility. *Linguistics in the Netherlands*, 23, **1**, 101-113 (2006).
9. Hull, D.A. and Grefenstette, G., Querying across languages: a dictionary-based approach to multilingual information retrieval. *Proc. 19th Annual Inter. ACM SIGIR Conf. on Research and Develop. in Infor. Retrieval*, New York, NY, USA. ACM, 49-57 (1996).
10. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., La Erty, J.D., Mercer, R.L. and Roossin, P.S., A statistical approach to machine translation. *Computational linguistics*, 16, **2**, 79-85 (1990).
11. Saers, M. and Wu, D., Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. *Proc. Third Workshop on Syntax and Structure in Statistical Translation*, Stroudsburg, PA, USA. Association for Computational Linguistics, 28-36 (2009).
12. Beloucif, M., Saers, M. and Wu, D., Improving word alignment for low resource languages using English monolingual SRL. *Proc. 26th Inter. Conf. on Computational Linguistics*, 51-60 (2016).
13. Echizen-ya, H., Araki, K. and Momouchi, Y., Automatic acquisition of bilingual rules for extraction of bilingual word pairs from parallel corpora. *Proc. ACL-SIGLEX Workshop on Deep Lexical Acquisition*, Stroudsburg, PA, USA, Association for Computational Linguistics, 87-96 (2005).
14. Rapp, R., Automatic identification of word translations from unrelated English and German corpora. *Proc. 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Stroudsburg, PA, USA, Association for Computational Linguistics, 519-526 (1999).
15. Fung, P., A statistical view on bilingual lexicon extraction. *Parallel Text Processing*, Springer, 219-236 (2000).
16. Haghighi, A., Liang, P., Berg-Kirkpatrick, T. and Klein, D., Learning bilingual lexicons from monolingual corpora. *Proc. ACL-08: HLT*, 771-779 (2008).
17. Irvine, A. and Callison-Burch, C., A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43, **2**, 273-310 (2017).
18. Nasution, A.H., Murakami, Y. and Ishida, T., Designing a collaborative process to create bilingual dictionaries of Indonesian ethnic languages. *Proc. Eleventh Inter. Conf. on Language Resources and Evaluation*, Miyazaki, Japan, 3397-3404 (2018).
19. Nasution, A.H., Murakami, Y. and Ishida, T., Plan optimization for creating bilingual dictionaries of low-resource languages. *Proc. Inter. Conf. on Culture and Computing (Culture and Computing)*, 35-41 (2017).
20. Nasution, A.H., Murakami, Y. and Ishida, T., Constraint-based bilingual lexicon induction for closely related languages. *Proc. Tenth Inter. Conf. on Language Resources and Evaluation*, Paris, France, 3291-3298 (2016).
21. Nasution, A.H., Murakami, Y. and Ishida, T., A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Infor. Process.*, 17, **2**, 9, 1-29 (2017).
22. Ishida, T. (Ed), *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer Publishing Company, Inc. (2011).

23. Ishida, T., Intercultural collaboration and support systems: a brief history. *Proc. Inter. Conf. on Principles and Practice of Multi-Agent Systems*, Springer, 3-19 (2016).
24. Nasution, A.H., Syafitri, N., Setiawan, P.R. and Suryani, D., Pivot-based hybrid machine translation to support multilingual communication. *Proc. Inter. Conf. on Culture and Computing (Culture and Computing)*, 147-148 (2017).
25. Nakaguchi, T., Takasaki, T., Pangaea, N., Otani, M. and Ishida, T., Combining human inputters and language services to provide multi-language support system for international symposiums. *Proc. Third Inter. Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies*, 28-35 (2016).
26. Linguistic Data Consortium. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Technical Report, Tech. Report (2005).