A Survey of Machine Translation Work in the Philippines: From 1998 to 2018

Nathaniel Oco naoco@national-u.edu.ph
Rachel Edita Roxas reoroxas@national-u.edu.ph
National University, Manila, 1008, Philippines

Abstract

In this paper, we present a survey covering the last 20 years of machine translation work in the Philippines. We detail the various approaches used and innovations applied. We also discuss the various mechanisms and support that keep the MT community thriving, as well as the challenges ahead.

1. Introduction

The Philippines is a country in Southeast Asia with 7,107 islands, and 187 languages – broken down as follows: 175 are indigenous, 8 are non-indigenous, and 4 are already extinct[1]. The official languages are: (1) Filipino, which was based on Tagalog and has 45 million L2 users[2]; and (2) English. Because the country is a gold mine for language data, there is already a thriving Natural Language Processing (NLP) community as evidenced by the existence of the following: the Computing Society of the Philippines – Special Interest Group on Natural Language Processing (CSP SIG-NLP)[3]; various institutions and research laboratories working on NLP[4,5]; venues to share ideas and knowledge like the National NLP Research Symposium (NNLPRS)[6]; and hosting of international conferences such as the 31st Pacific Asia Conference on Language, Information and Computation or PACLIC 31 (2017)[7].

One work (Raga, 2016) looked at the progress of NLP research in the country by reviewing 12 editions of NNLPRS. Our work differs in that we are focused on machine translation and we also looked at other venues. With NNLPRS as starting point, we branched out to cover the references the authors cited, other conference proceedings, journal issues, and works by members of CSP SIG-NLP. We took note of the approaches used, data size, and innovations applied.

The rest of the paper is organized as follows: in section 2, we present the evolution of machine translation in the Philippines by presenting the various approaches used over time; in section 3, we discuss the different innovations applied from data collection to evaluation; we tackle the challenges in section 4; and we conclude our work in section 5.

[1] Data according to Ethnologue: https://www.ethnologue.com/country/PH
[2] Data according to Ethnologue: https://www.ethnologue.com/language/fil
[3] Computing Society of the Philippines: http://csp.org.ph/
[4] National University's Research and Innovation Office: http://www.national-u.edu.ph/?page_id=44
[5] De La Salle University's Center for Language Technologies: http://www.dlsu.edu.ph/research/centers/adric/nlp/
[6] Website of the recently concluded student research workshop organized by CSP SIG-NLP: https://sites.google.com/bicol-u.edu.ph/14nnlprs-pre-conference/home
[7] PACLIC 31 (2017): http://paclic31.national-u.edu.ph/

2. Approaches

Machine translation started in the late '90s covering the two official languages (Roxas et al., 1999): Filipino/Tagalog and English. It later on included other Philippines languages[8] such as Cebuano (Yara, 2007), Kankanaey (Ananayo et al., 2011), Maranao (Dimalen et al., 2009), Hiligaynon (Macabante et al., 2017), Ilocano (Miguel and Dy, 2008; Bautista et al., 2015; Lazaro et al., 2017), and Bicol (Fernandez et al., 2018). Applications of machine translation since then can be grouped into three: (1) in tourism (Lazaro et al., 2017); (2) in translating informational materials such as books for mother tongue-based – multilingual education or MTB-MLE (Oco et al., 2016); and (3) in humanitarian technologies, for example to assist policy makers make sense of community input (Octaviano et al., 2018). We've seen that early works only tackled declarative and imperative statements but the advent of statistical machine translation (Nocon et al., 2014) paved the way to also include interrogative statements. We have observed that all serious research undertaking has been supported in part by government funding. It started with transfer-based approaches and succeeding projects have seen rule-based, corpus-based, statistical, and deep learning approaches. In the succeeding text, we discuss these projects and the approaches used, and direction.

2.1. Transfer-based approaches

Machine translation in the Philippines traces its early roots to transfer-based approaches. One such project is IsaWika! (Roxas et al., 1999), an English-Tagalog machine translator for declarative and imperative sentences, that used an augmented transition network and a dictionary size of less than 10,000 entries. The project's second phase started in 1998 and was funded by the Department of Science and Technology – Philippine Council for Advanced Science and Technology Research and Development (DOST-PCASTRD). This was immediately followed by a project (Borra, 1999) which explored lexical functional grammar or LFG as the grammar formalism. The f-structure and c-structures also showed promise in identifying translation errors. LFG would be a staple in machine translation projects with XLE parser (Frank et al., 1998) as the core. One project (Borra et al., 2007) used a transfer dictionary with 2,000 parallel word pairs while another project (Cada et al., 2012) used a bootstrapping technique to develop a larger parallel corpus from earlier works. Recent developments include the use of a natural language generator called Linguist's Assistant (Allman et al., 2014) to translate religious text[9]. It is being used to build lexicons and grammars in Filipino, Ayta Mag-indi, and Botolan languages, and can be used towards developing materials for mother tongue-based – multilingual education or MTB-MLE (Oco et al., 2016), a form of education where children's mother tongue are used as the primary mode of teaching until primary school. In all transfer-based projects, we have noticed that the corpus size is limited and vocabulary is only within the scope of available resources.

2.2. Corpus-based

Seeing the limitations of manually creating rules in a transfer-based approach, various corpora were soon utilized. After IsaWika!, DOST-PCASTRD funded a hybrid English-Filipino machine translation system from 2005 to 2008 (Roxas, 2006; Roxas et al., 2008). It combined both transfer-based and corpus-based approaches.

[8] Both Kankanaey and Maranao are considered indigenous languages

[9] A version of Linguist's Assistant called "The Bible Translator's Assistant" is being used to translate books of the Bible to low-resource Philippine languages. Website: http://www.thebibletranslatorsassistant.org/

The transfer-based approach used an LFG formalism while the corpus-based approaches extracted patterns (Alcantara et al., 2006) from a large document and stored them in templates (Go et al., 2007). The project used a parallel corpus with 207,000 Filipino words and a dictionary with 4,000 words. For hybrid systems, the challenge is integrating results from multiple machine translators. One solution is to develop a module that can provide translation scores.

2.3. Statistical

The Network-based ASEAN Languages Translation Public Service or ASEANMT saw the introduction of statistical approaches. It aims to "*build a practical network-based service on ASEAN languages text translation in the tourism domain. ASEAN languages resources and knowledge of the translation technology are availably shared among ASEAN member states and other countries*"[10]. It is supported by the Association of Southeast Asian Nations Committee on Science and Technology (ASEAN COST). The Center for Language Technologies at De La Salle University represented the country in this project with funding from the Commission on Higher Education for the counterpart system (Ilao et al., 2015; Nocon et al., 2014). Moses engine[11] was used with covering 20,000 sentence pairs on the tourism domain and at least 100,000 thousand sentence pairs derived from Wikipedia articles and manually translated. A demo version is available online[12].

2.4. Directions

We see the direction of machine translation to be heading towards deep learning because of the availability of approaches to automatically build parallel corpora. One work (Tacorda et al., 2017), also supported by government funding[13], utilized RNN with 100,000 pairs of sentences and integrated byte pair encoding (BPE) to reduce out-of-vocabulary errors (OOVs). BPE works by segmenting a token into identifiable sequences. This allowed for tokens not present in the training data to be recognized if its root and affixes have been identified through BPE. The danger is with false positives: character sequences part of the root can be identified falsely as an affix.

Aside from deep learning, recent trends in machine translation focused on its application in humanitarian technologies. As example, one project (Octaviano et al., 2018) is involved in eParticipation, specifically in cross-lingual topic modeling – translating community responses and generating topic models through LDA – to make sense of community inputs. Qualitative evaluation showed cross-lingual topic modeling generated more coherent topic models. Another work (Fernandez et al., 2018) aims at assisting non-linguists in translating questions for survey use.

3. Innovations

Aside from BPE, other innovations to reduce OOVs include the use of domain adaptation techniques (Lazaro et al., 2017) by filtering the training data, and allowing users to provide correction through feedback (Ang et al., 2015). Other projects addressed OOVs by increasing the training data through automatic means. One work (Dimalen and Roxas), crawled the web

[10] Website: http://aseanmt.org/

[11] Website: http://www.statmt.org/moses/

[12] Demo version: http://www.aseanmt.org/mt/

[13] Supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (no. IIID-2015-07)

and automatically identified the language through a trigram ranking approach. Odds-ratio was applied since closely related languages yield to lower recall rates. Other researchers have attempted to find bilingual pairs of terms (Lat et al., 2006) and sentences (Tabaranza et al., 2016) in comparable corpora. There are also attempts to gamify manual translation (Ilao et al., 2016) in role-playing games: if the user wants to earn more credit points, he/she can translate phrases and there's an automatic scoring mechanism that rewards the user after a given time frame. There are also those (Octaviano et al., 2018) that apply spell checking and language identification as pre-processing step to clean the data. To assist translators, one work (Oco and Borra, 2011) utilized Transifex in localizing labels and instructions. Another allowed linguists to provide semantic representations (Allman et al., 2014).

As for evaluation, the ASEAN MT asked manual annotators to evaluate machine translation output and provide a rating from 1 to 5, with the highest having semantic equivalence with the manually translated one. Another (Allman et al., 2014) asked students to read manually translated and automatically translated materials and an assessment task was given. Those who were given the automatically translated material as reference scored higher than those who were given manually translated materials.

4. Challenges

Aside from free-word order, there are other challenges that make translation work in the Philippines interesting:

- Verbs have both tense and focus (Ramos and Cena, 1990).
- Affixes can be divided into prefix, infix, suffix, circumfixation, and there is also affix reduplication (Schachter and Otanes, 1972).
- Plurality exists in pronouns, modifiers, and verbs (Ramos, 1971; Cubar and Cubar, 1994; Kroeger, 1993).

5. Conclusion

We have surveyed projects covering the last 20 years of machine translation work in the country. We have observed that funding and support from the government combined with venues that allow the flow and sharing of knowledge enabled researchers to advance the growth of the field. The lack of available resources provided researchers problems to work on and for innovations to surface. Through various means, we noted that researchers are able to be constantly updated on recent trends. Most of the works we presented in this paper focused only on text and it highlights that there are still room for speech to speech translation.

Acknowledgement

This work is supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (no. IIID-2015-07).

References

Alcantara,D.L., Hong, B.A., Perez, A., Tan, L. and Tan, M.W. (2006). Rule Extraction Applied in Language Translation. In *Proceedings of the 3rd Natural Language Processing Research Symposium*, pages 19–22, Manila, Philippines.

Allman, T., Beale, S. and Denton, R. (2014). Toward an Optimal Multilingual Natural Language Generator: Deep Source Analysis and Shallow Target Analysis. *Philippine Computing Journal*, 9(1): 55–63.

Allman, T. (2015). Linguist's Assistant: Gleaning a Tagalog Lexicon and Grammar from a Small, Light-ly Annotated Corpus. In *Proceedings of the 11th Natural Language Processing Research Symposium*, page 1, Manila, Philippines.

Ananayo, J., Cayaos, J.D. and Rosal, F.G. (2011). Translation Algorithm: English to Kankanaey. In *Proceedings of the 8th Natural Language Processing Research Symposium*, pages 11–16, Manila, Philippines.

Ang, J., Chan, M.R., Genato, J.P., Uy, J. and Ilao, J. (2015). Development of a Filipino-to-English Bidirectional Statistical Machine Translation System that dynamically updates via user feedback. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Bautista, J., Bayla, C., Fianza, K., Mamis, D., Tangangco, J., Yango, J. and Miguel, D. (2019). Bidirectional Ilocano-English Language Translator Using Customized Moses Statistical Machine Translation System (SMTS). In *Proceedings of the 11th Natural Language Processing Research Symposium*, pages 18–25, Manila, Philippines.

Borra, A. (1999). A Transfer-Based Engine for an English to Filipino Machine Translation Software. MS Thesis, University of the Philippines Los Baños.

Borra, A. Chan, E.A., Lim, C.I., Tan, R.B. and Tong, M.C. (2007). LFG-Based Machine Translation Engine for English and Filipino. In *Proceedings of the 4th Natural Language Processing Research Symposium*, pages 36–42, Manila, Philippines.

Cada, D.R., Chan, F.A., Chen, H.Z. and Tan, A.E. (2012). Bootstrapping a Tagalog LFG F-structure Bank. Undergraduate Thesis, De La Salle University.

Cubar, E. and Cubar, N. (1994). *Writing Filipino Grammar: Traditions and Trends*. New Day Publishers.

Dimalen, D.M., Dimalen, E., Pangandaman, M. and Wade, J. (2009). MELT: Towards Buidling an Indigenous MT System in Meanao to English Language. In *Proceedings of the 6th Natural Language Processing Research Symposium*, pages 34–37, Manila, Philippines.

Dimalen, D.M. and Roxas, R. (2007). AutoCor: A Query Based Automatic Acquisition of Corpora of Closely-related Languages. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 146–154, Seoul, South Korea.

Fernandez, J.D., Jadie, J., Lim, C.K., Zuniega, J. and Llovido, J. (2018). Bi-directional Bikol-English Statistical Machine Translator. Presented at the *14th National Natural Language Processing Research Symposium Pre-Conference Activity: Student Research Workshop*, Legazpi, Philippines.

Frank, A., King, T.H., Kuhn, J. and Maxwell, J. (1998). Optimality Theory style constraint ranking in large-scale LFG grammars. In *Proceedings of the 3rd LFG Conference*, Brisbane, Australia.

Ilao, J., Roxas, R., Sison-Buban, R., Cheng, C., See, S. and Regalado, R.V. (2016). Philippine Component of the ASEAN Machine Translation Project. Paper presented at the *12th Natural Language Processing Research Symposium*, Dumaguete, Philippines.

Kroeger, P. (1993). *Phrase Structure and Grammatical Relations in Tagalog*. CSLI Publications.

Lat, J.O., Ng, S.T., Sze, K., Yu, G.D. and Lim, N.R. (2006). Lexicon Acquisition for the English and Filipino Language. In *Proceedings of the 3rd Natural Language Processing Research Symposium*, pages 49–54, Manila, Philippines.

Lazaro, A.N., Oco, N. and Roxas, R.E. (2017). Developing a Bidirectional Ilocano-English Translator for the Travel Domain: Using Domain Adaptation Techniques on Religious Parallel Corpora. Presented at the *11th International Conference of the Asian Association for Lexicography*, Guangzhou, China.

Macabante, D.G., Tambanillo, J.C. Dela Cruz, A. Ellema, N., Octaviano, M. Rodriguez, R. and Roxas, R.E. (2017). Bi-directional English-Hiligaynon statistical machine translation. In *Proceedings of TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 2852–2853, Penang, Malaysia.

Miguel, D. and Dy, M.C. (2008). ANGLOCANO: an Ilocano to English Machine Translation System. In *Proceedings of the 5th Natural Language Processing Research Symposium*, pages 85–92, Manila, Philippines.

Nocon, N., Oco, N., Ilao, J. and Roxas, R E. (2014). Philippine Component of the Network-based ASEAN Language Translation Public Service. In *Proceedings of the 7th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management*, Puerto Princesa City, Philippines.

Oco, N. and Borra, A. (2011). A Grammar Checker for Tagalog using LanguageTool. In *Proceedings of the 9th Workshop on Asian Language Resource Collocated with IJCNLP 2011*, pages 2–9, Chiang Mai, Thailand.

Oco, N., Syliongka, L.R., Allman, T. and Roxas, R.E. (2016). Resources for Philippine Languages: Collection, Annotation, and Modeling. In *Proceedings of the 30th Pacific Asia Conference on Language, Information, and Computation*, pages 433–438, Seoul, South Korea.

Octaviano, M., Dela Cruz, A., Oco, N. and Roxas, R.E. (2018). Cross-Lingual Topic Discovery. Presented at the *18th Philippine Computing Science Congress*, Cagayan de Oro, Philippines.

Raga, R. (2016). Reflections on the Awareness and Progress of Natural Language Processing (NLP) Research in the Philippines. *Philippine Computing Journal*, 11(1):1–9.

Ramos, T. (1971). *Makabagong Balarila ng Pilipino*. Rex Book Store.

Ramos, T. and Cena, R. (1990). *Modern Tagalog: Grammatical Explanations and Exercises for Non-native Speakers*. University of Hawaii Press.

Roxas, R.E., Sanchez, W. and Buenaventura, M. (1999). Machine Translation from English to Filipino: Second Phase. Report submitted to the Department of Science and Technology – Philippine Council for Advanced Science and Technology Research and Development.

Roxas, R.E. (2006). A Hybrid English-Filipino Machine Translation System. In *Proceedings of the 3rd Natural Language Processing Research Symposium*, pages 1–4, Manila, Philippines.

Roxas, R.E., Borra, A., Cheng, C., Lim, N.R., Ong, E.C. and Tan, M.W. (2008). Building language resources for a Multi-Engine English-Filipino machine translation system. *Language Resources and Evaluation*, 42:183–195.

Schachter, P. and Otanes, F. (1972). *Tagalog Reference Grammar*. University of California Press.

Tabaranza, Z.L., Bureros, L. and Roxas, R. (2016). English-Cebuano Parallel Language Resource for Statistical Machine Translation System. In *Proceedings of the 11th International Symposium on Natural Language Processing*, Ayutthaya, Thailand.

Tacorda, A.J., Ignacio, M.J., Oco, N. and Roxas, R.E. (2017). Controlling Byte Pair Encoding for Neural Machine Translation. In *Proceedings of the 21st International Conference on Asian Language Processing*, Singapore, Singapore.

Yara, J. (2007). A Tagalog-to-Cebuano Affix-Transfer-Based Machine Translator. In *Proceedings of the 4th Natural Language Processing Research Symposium*, pages 32–35, Manila, Philippines.