Resources for Philippine Languages: Collection, Annotation, and Modeling

Nathaniel Oco[a], Leif Romeritch Syliongka[a], Tod Allman[b], Rachel Edita Roxas[a]
[a]National University
551 M.F. Jhocson St., Sampaloc, Manila, PH 1008
[b]Graduate Institute of Applied Linguistics
7500 W. Camp Wisdom Rd., Dallas, TX 75236
{nathanoco,lairusi,todallman,rachel_roxas2001}@yahoo.com

Abstract

In this paper, we present our collective effort to gather, annotate, and model various language resources for use in different research projects. This includes those that are available online such as tweets, Wikipedia articles, game chat, online radio, and religious text. The different applications, issues and directions are also discussed in the paper. Future works include developing a language web service. A subset of the resources will be made temporarily available online at: http://bit.ly/1MpcFoT.

1 Introduction

The Philippines is a country in Southeast Asia composed of 7,107 islands and 187 listed individual languages. Among these, 41 are listed as institutional, 73 are developing, 45 are vigorous, 13 are in trouble, 11 are dying, and 4 are already extinct[1]. These numbers highlight that there is a pressing need for a databank on Philippine languages. As highlighted in literature (Dita et al., 2009; Oco and Roxas, 2012), even those with high number of native speakers have limited available corpora. Towards addressing this scenario, we describe in this paper the collection, annotation, and modeling of various language resources.

The paper's structure is as follows: section 2 discusses initiatives in the country and the various language resources we collected; section 3 discusses annotation and documentation efforts; section 4 discusses language modeling; and we conclude our work in section 5.

2 Collection

Research works in language studies in the Philippines – particularly in language documentation and in corpus building – often involve one or a combination of the following: "(1) residing in the place where the language is spoken, (2) working with a native speaker, or (3) using printed or published material" (Dita and Roxas, 2011). Among these, working with resources available is the most feasible option given ordinary circumstances. Following this consideration, the Philippines as a developing country is making its way towards a digital age, which highlights – as Jenkins (1998) would put it – a "technological culture of computers". Organizations and educational institutions are making resources available in the Internet.

In the Philippines, documenting languages and making the resources public had been realized even before the turn of the millennium. For our collection initiatives, we derive inspiration from previous works. One of the projects is IsaWika (Roxas and Borra, 2000). It is an English-Filipino machine translator developed in 1999 that translates simple declarative sentences using the augmented transition network. Years after, the development of the Philippine component of the International Corpus of English or ICE-PHI (Bautista, 2004) started. It contains one million words of written and spoken Philippine English.

[1] Ethnologue Philippine language status profile for the Philippines: http://www.ethnologue.com/country/PH

Source	Language	Number of Articles
Inquirer	English	547
Manila Bulletin	English	1,333
Pang-Masa	Filipino	576
Pilipino Star Ngayon	Filipino	1,013
Rappler	English	779
The Philippine Star	English	1,011
Total		5,259

Table 1. Number of news articles collected

The written component contains non-printed texts such as non-professional writing and correspondence; and printed texts such as academic writing, reportage, instructional writing, persuasive and creative writing. On the other hand, the transcribed texts contain dialogues and monologues. One feature of the corpus is manual annotation – there are markup symbols to indicate the part-of-speech, the foreign and indigenous words, and paralinguistic devices.

The advent of ICE-PHI inspired the development of PALITO (Dita et al., 2009). An online corpus developed for purposes of pursuing various linguistic agendas, PALITO is a repository for religious and literary texts written in eight Philippine languages – Bikol, Cebuano, Hiligaynon, Ilocano, Kapampangan, Pangasinense, Tagalog, and Waray – and has a total word count of two million words. Aside from written texts, another component is the Filipino sign language (FSL) videos. The signs were illustrated in the form of actions and the videos cover the alphabet, number system, basic terms and expressions, and discourse. The 118 FSL videos have a combined file size of 224 million bytes.

The size of both ICE-PHI and PALITO highlights one limitation if collection was done manually – the low number of resources. In our initiative, we address this through automatic means. Current statistics put the number of Internet users to 44% [2], which makes social media and other online forms as viable sources of data, and one popular form is Twitter, a social networking site.

2.1 Tweets

For the current work, we continued the automatic collection of tweets started in a previous project (Oco et al., 2014b). A program was used to collect these online. A tweet is a short 140-character message delivered in Twitter. The program uses Twitter 4J[3] Java library and stores the following information in a database: Tweet ID; the tweet; date and time the tweet was sent; and geolocation where the tweet was sent.

The collection started last February 17, 2013 and a total of approximately 50 million tweets have been collected as of this writing. As tweets are considered an online chronicle of events and a repository of human opinion, they have been used to study Filipino voting behavior (Pablo et al., 2014) and analyze disasters such as the typhoon Haiyan (Soriano et al., 2016). In tandem with classification techniques such as sentiment analysis, tweets can be used in prediction and to help policy makers make informed decisions. It should be noted that we only collected those that are publicly available and those whose location is set, following ethical standards.

2.2 News Articles

Aside from tweets, we also collected news articles as they also represent actual language usage. Also a continuation of a previous project (Oco et al., 2014b), we are collecting from four news agencies (shown in Table 1). Calibre, an open source e-book management system [4] was used. As of this writing, a total of 5,259 articles have been collected, which are in .txt, .epub, and .pdf formats.

News articles provide a clear usage of the language. In a culturomics study (Ilao et al., 2011), cultural trends were studied using articles collected from more than ten Philippine tabloids and a computational model for language development was built.

2.3 Game Chat

An untapped source of valuable information is massively multiplayer online role-playing games (or MMORPG). They provide a venue for people to interact virtually. One of the popular MMORPG in the Philippines is Ragnarok. Also a continuation of a previous project (Oco et al., 2014b), we are collecting chat logs from this game using OpenKore [5] – a client and bot program made specifically for Ragnarok. Chats are classified into four: (1) A private chat [PM], which is a message sent to the character and can

[2] https://telehealth.ph/2015/03/26/internet-social-media-and-mobile-use-of-filipinos-in-2015/

[3] http://twitter4j.org/en/index.html

[4] http://calibre-ebook.com/

[5] http://www.openkore.com/index.php/Main_Page

only be read by the recipient of the message; (2) a public chat [C], which is a message sent by nearby characters and can be read by other nearby players; and (3) a shout [GM] and [S], which are game-wide message and can be read by everyone.

The chat logs can be used as training data in text normalization (Nocon et al., 2014b) and in cyber bullying detection (Cheng and Ng, 2016). The current log has 1,166,352 lines.

2.4 Wikipedia

Inspired by a previous study that collected English and Tagalog Wikipedia articles (Oco et al., 2014b), we also collected Wikipedia articles in other Philippine languages because of its popularity and availability. Embodiment of majority of the society, it is a multilingual Internet encyclopedia that is collaboratively edited by volunteers. Entire Wikipedias are publicly available through XML dumps and editions in Philippine languages exist. As a form of cleaning, we used an XML to text converter[6] to extract entries from XML dumps of the following languages: Bikol, Chavacano Zamboangeño, Ilokano, Kapampangan, Pangasinense, and Tagalog. Table 2 shows a sample text in the XML dump and its converted version, while Table 3 shows the word count of the converted text. A total of 7,304,254 words were collected. There is also an ongoing work to collect articles from the Cebuano, Hiligaynon, and Waray Wikipedias. Earlier projects utilized the Tagalog Wikipedia for purposes of code-switching point detection (Oco and Roxas, 2012). Another work (Syliongka and Oco, 2014) utilized Wikipedia articles as training data for named entity extraction.

XML dump	Extracted Text
<text xml:space="preserve">{{year nav|{{PAGENAME}}}} Ang '''2005''' ay isang karaniwang taon na nagsisimula sa [[Sabado]] ayon sa [[Gregorian calendar]]. Ito ay hinirang na	Ang 2005 ay isang karaniwang taon na nagsisimula sa Sabado ayon sa Gregorian calendar. Ito ay hinirang na

Table 2. Sample XML dump

ISO Code	Word Count
bik	466,096
cbk	283,798
ilo	842,373
pag	127,492
pam	628,948
tgl	4,955,547
Total	7,304,254

Table 3. Summary of Wikipedia articles

ISO Code	**Corpus Size (Words)**
ceb	1,069,713
hil	291,370
ilo	1,003,392
tgl	1,104,035
Total	**3,468,510**

Table 4. Corpus size of Bible editions

2.5 Bible Editions

A number of religious organizations have put efforts to translate the Bible and made them available online in an effort to promote biblical teachings. One of these organizations is the Jehovah's Witness, which has more than 100,000 active members in the Philippines. It provides ePub versions of several Bible editions in their website[7]. To collect these, we used Calibre to convert the files to text files. The size of each edition is detailed in Table 4. The collection has a total size of 3,468,510 words. Bible editions have been used as training data in translation studies, language identification (Oco et al., 2013a; Oco et al., 2014a; Octaviano et al., 2015), and language clustering (Oco et al., 2013b).

2.6 Online Radio

Deviating from the usual text collection, we have also decided to start collecting audio files. A number of radio stations in the Philippines are already airing online. We used Screamer Radio[8] to record airings from radio stations. Figure 1 shows a sample screenshot of the program. More than 5,400 hours of music, commercials, and commentaries have been recorded in stereo format from three radio stations. The details are shown in Table 5. Screamer radio provides direct MP3 audio stream saving at 31kbps. These audio files could be used as training data for speech synthesis and automatic speech recognition (ASR). One study (Laguna and

[6] An XML to text converter by Evan Jones: http://www.evanjones.ca/software/wikipedia2text.html

[7] http://www.jw.org/en/publications/bible/

[8] http://www.screamer-radio.com/

Guevara, 2014), developed an audio language identification system for different Philippine languages.

3 Annotation and Documentation

Another corpus being developed consists of lexicons, grammars, and texts that are translated by a natural language generator (NLG) called Linguist's Assistant (LA) (Beale and Allman, 2011). LA is being used to build lexicons and grammars for many of the languages in the Philippines, and that data is being applied to thoroughly annotated semantic representations in order to generate initial draft translations of various educational, religious, and community development texts. LA has successfully been used to document many languages from a variety of language families, and it has produced high quality initial draft translations of a variety of texts in those languages (Allman et al., 2014). A new technique developed specifically for the Malayo-Polynesian languages of the Philippines was recently implemented (Allman, 2014), and an extensive Tagalog lexicon and grammar were developed. LA is now able to produce initial draft translations of a variety of texts in Tagalog. Experiments indicate that mother-tongue translators are able to edit those drafts into publishable form in approximately one fourth the time required for manual translation. The Tagalog grammar and lexicon are currently being expanded so that LA can produce initial draft translations of a wider variety of texts. Additionally, the Tagalog grammar and lexicon are being modified to accommodate another Malayo-Polynesian language named Ayta Mag-Indi. Because Tagalog and Ayta Mag-Indi are structurally very similar, the process of modifying the Tagalog grammar to accommodate Ayta Mag-Indi is proceeding very quickly. After only six meetings with an Ayta speaker, LA was able to produce an initial draft translation of a simple story in Ayta Mag-Indi. That same story required approximately 38 meetings with the Tagalog mother-tongue speaker. The work to expand the Ayta Mag-Indi grammar and lexicon will continue so that more texts can be translated. After the Ayta Mag-Indi work has been completed, the same process will be repeated with multiple Malayo-Polynesian languages. There are many potential uses for the resulting data and corpora. A few of these usages are detailed in the following sections.

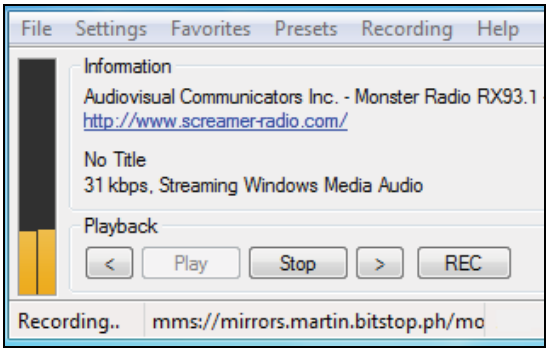

Figure 1. A screenshot of Screamer Radio

Description	Love Radio	Monster Radio	Yes FM
Frequency	90.7 MHz	93.1 MHz	101.1 MHz
Power	25 kW	25 kW	25 kW
Total File Size	19.7 GB	32.5 GB	
Total Length (hours)	2,532	1,869	1,075

Table 5. Details about the Radio Stations

3.1 MTBMLE

The Mother-Tongue Based Multilingual Education (MTBMLE) initiative mandates that many educational books and instructional documents be translated into all of the Philippine languages. After a lexicon and grammar have been developed for a particular language, LA is able to generate a translation of any document that has been converted into a semantic representation. Therefore LA could potentially facilitate the implementation of MTBMLE.

3.2 ASEAN-MT

The ASEAN Machine Translation project or ASEAN-MT (Nocon et al., 2014a) uses standard English as its interlingua. However, a problem will arise because English is impoverished in many areas (e.g., its pronominal system, its deictic and article systems, its tense system, etc.). The Tagalog texts produced by this project could be aligned with the associated semantic representations, and then the stochastic trainer for the Tagalog component of ASEAN-MT could be used to produce enriched English as the interlingua. Using enriched English as the interlingua will significantly improve the quality of the translations produced in the other ASEAN languages.

3.3 Linguistic Research

Having computational lexicons and grammars in a standardized format for many of the Philippine languages could prove invaluable to linguists, lexicographers, and translators throughout the country. Additionally the data could be used to determine language affinities, supplement cross-linguistic research, and bolster typological studies.

A tool that can be used by linguists for describing morphology, syntax, and semantics is beneficial (Beale et al., 2005). Anthropologists have noted that when a language is documented and texts are translated into the language, the speakers of the language are often motivated to preserve their language and expand its use.

4 Modeling

A number of the resources were modeled in terms of word n-grams and character n-grams. A word n-gram is n-slices of a sentence while a character n-gram is n-slices of a word. As an example, the list of 3-grams that can be generated from the word "language" are: {_la, lan, ang, ngu, gua, uag, age, ge_}. These language models provide information on frequently occurring words and phrases. The advent of data-centric computing made it possible for models to be used as training data in various tasks (Legaspi et al., 2008; Gavrila and Vertan, 2011). A number of text documents written in various languages have been used as training data in language identification (Oco et al., 2013a) and language clustering (Oco et al., 2013b).

5 Conclusion

In this paper, we presented our collective effort to collect, annotate, document, and model various language resources to address the pressing need for a databank on Philippine languages. We also detailed the different applications, issues, and directions.

The work can be extended by considering automatic annotation and making the resources available online as a language web service.

Acknowledgment

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

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

Allman, T. 2014. Linguist's Assistant: Gleaning Malayo-Polynesian Grammars from Small, Lightly Annotated Corpora. Paper presented at the 12th Philippine Linguistics Congress.

Bautista, M.L. 2004. An Overview of the Philippine Component of the International Corpus of English (ICE-PHI). Asian English, 7(2), pp. 8-26.

Beale, S., S. Nirenburg, M. McShane, and T. Allman. 2005. Document Authoring the Bible and for Minority Language Translation. Proceedings of MT Summit.

Beale, S. and T. Allman. 2011. Linguist's Assistant: A Resource for Linguists. Proceedings of the 9th Workshop on Asian Language Resources, pp. 41-49.

Cheng, C. and L. Anson Ng. 2016. Automated Role Detection in Cyberbullying Incidents. Proceedings of the 16th Philippine Computing Science Congress, pp. 85-92.

Dita, S., R.E. Roxas, and P. Inventado. 2009. Building Online Corpora of Philippine Languages. Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, pp. 646-653.

Dita, S. and R.E. Roxas. 2011. Philippine Languages Online Corpora: Status, Issues, and Prospects. Proceedings of the 9th Workshop on Asian Language Resources, pp. 59-62.

Gavrila, M. and C. Vertan. 2011. Training Data in Statistical Machine Translation: The More the Better. Proceedings of Recent Advances in Natural Language Processing, pp. 551-556.

Ilao, J., R.C. Guevara, V. Llenaresas, E.A. Narvaez, and J. Peregrino. 2011. Bantay-Wika: Towards a Better Understanding of the Dynamics of Filipino Culture and Linguistic Change. Proceedings of the 9th Workshop on Asian Language Resources, pp. 10-17.

Jenkins, H. 1998. The Poachers and the Stormtrooper: Popular Culture in the Digital Age. Red Rock Eaters News.

Laguna, A.F.B. and R.C.L. Guevara. 2014. Experiments on Automatic Language Identification for Philippine Languages using Acoustic Gaussian Mixture Models. Proceedings of IEEE TENSYMP.

Legaspi, R., S. Kurihara, K. Fukui, K. Moriyama, and M. Numao. 2008. Proceedings of the 5th

International Conference on Information Technology and Applications, pp. 88-93.

Nocon, N., G. Cuevas, D. Magat, P. Suministrado, and C. Cheng. 2014a. NormAPI: An API for normalizing Filipino shortcut texts. Proceedings of the International Conference on Asian Language Processing.

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

Oco, N. and R.E. Roxas. 2012. Pattern Matching Refinements to Dictionary-Based Code-Switching Point Detection. Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation, pp. 229-236.

Oco, N., J. Ilao, R.E. Roxas, and L.R. Syliongka. 2013a. Measuring Language Similarity using Trigrams: Limitations of Language Identification. Proceedings of the 3rd International Conference on Recent Trends in Information Technology.

Oco, N., L.R. Syliongka, J. Ilao, and R.E. Roxas, 2013b. Dice's Coefficient on Trigram Profiles as Metric for Language Similarity. Proceedings of the 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE).

Oco, N., L.R. Syliongka, J. Ilao, and R.E. Roxas. 2014a. N-gram based Language Identification and Rule-based Grammar Checking. Proceedings of the 14th Philippine Computing Science Congress, pp. 244-250.

Oco, N., R. Sison-Buban, L.R. Syliongka, R.E. Roxas, and J. Ilao. 2014b. Ang Paggamit ng Trigram Ranking Bilang Panukat sa Pagkakahalintulad at Pagkakapangkat ng mga Wika [Trigram Ranking: Metric for Language Similarity and Clustering]. Malay, 26(2), pp 53-68.

Octaviano Jr., M., R. Fajutagana, C.M.L., J.D. Miñon, J.-A. Morano, R.C. Tinoco, and N. Oco. 2015. The use of Trigram Models in Classifying and Clustering different Philippine Languages. Proceedings of the 10th International Conference on Knowledge Information and Creativity Support Systems, pp. 546-552.

Pablo, Z.C., N. Oco, M.D.G. Roldan, C. Cheng, and R.E. Roxas. 2014. Toward an enriched understanding of factors influencing Filipino behavior during elections through the analysis of Twitter data. Philippine Political Science Journal, 35(2), pp. 203-224.

Roxas, R. and A. Borra. 2000. Computational Linguistics Research on Philippine Languages. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

Soriano, C.R., M.D.G. Roldan, C. Cheng, and N. Oco. 2016. Social media and civic engagement during calamities: The case of Twitter use during typhoon Yolanda. Philippine Political Science Journal, 37(1), pp. 6-25.

Syliongka, L.R. and N. Oco. 2014. Using Language Modeling and Data Association to Perform Named Entity Recognition. Proceedings of the 10th National Natural Language Processing Research Symposium, pp. 115-119.