Practical Applications of Human Language Technology: the Philippine Experience

Rachel Edita O. Roxas

College of Computer Studies
De La Salle University-Manila
2401 Taft Avenue, Manila Philippines
Tel: (63) (2) (536-0276/7)
Fax: (63) (2) (536-0278)

rachel.roxas@delasalle.ph

ABSTRACT

In this paper, I present the practical applications of human language technology researches that we have conducted in the country. These applications include fields such as in history, culture and the arts, education, law and business. The discussions show that the development of applications for HLT in the country provides a very rich and diverse platform for further research and study, opening up many emerging areas and fields of study.

Keywords

Applications of Natural Language Processing, Corpora

1. CULTURE AND THE ARTS

The Philippines is an archipelago of more than 7,100 islands, with over 171 languages (according to the SIL http://www.sil.org/asia/philippines/). The country has a rich culture as evidenced by the diversity of its languages.

As part of our culture, language has a key role. An online repository of documents on Philippine languages is provided through the Palito website [11]. Initial languages in the corpora include Tagalog, Cebuano, Ilocano and Hiligaynon with manually-collected 250,000 words each (in text), and the Filipino sign language with 7,000 signs (in video). The users of the system can also view specific documents using its internal search engine and most importantly use different tools for linguistic research, such as document search, word count, and word concordancer. The concordancer for a particular word in the corpus pertains to a list of the occurrences of the specified word in a document or a group of documents, showing its immediate context.

The Filipino Sign Language (FSL) component of the Philippine language corpora includes signs and discourse in video, which are edited, glossed and transcribed. Video editing cuts the video for final rendering, glossing allows association of sign to particular words, and transcription allows viewing of textual equivalents of the signed videos. The FSL corpus videos can be viewed, together with their transcriptions and some with gloss (Figure 1).

The future goal of this project is to be able to collect the corpora for other Philippine languages, and to develop the tools for linguistic studies and researches on these languages.

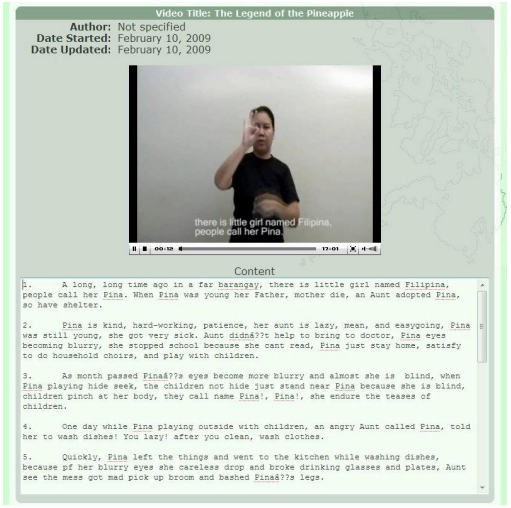

Figure 1. Sample Video with Gloss and Transcription Viewed in Palito

To augment these manually collected corpora on Philippine languages, automatic tools are also being developed to handle automatic retrieval. This includes studies such as those of Dimalen & Roxas [10], which provided a methodology to automatically distinguish the differences of closely-related languages such as Philippine languages.

Another very interesting aspect of language documentation is the tracing of the historical development of our language through the centuries. This entails an unexplored area that involves the collection of these historical documents and the corresponding historical (and possibly, automatic) tracing of these documents [15]. An interesting piece of historical information is in *Doctrina Christiana*, the first ever published work in the country in 1593 which shows the translation of religious material in the local Philippine script, the Alibata, old Tagalog, and Spanish. Current digitalization efforts include scanned pages of the document. A sample page from *Doctrina Christiana* is shown in Figure 2, courtesy of the University of Santo Tomas Library.

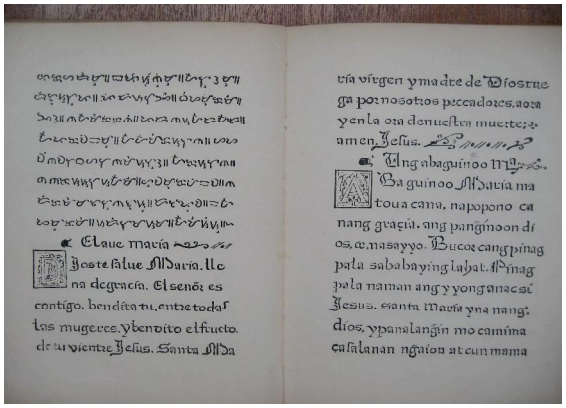

Figure 2. A Page from Doctrina Christiana

2. EDUCATIONAL APPLICATIONS

Several language applications have also been developed that are specifically prepared as an aide to the teaching of Philippine languages. Although most of these works focus on the English language, explorations are encouraged to consider the Philippine languages towards the fulfillment of the mother-tongue multi-lingual education advocated by various sectors in the country. Language applications for teaching include language learning, reading comprehension and composition writing; the discussions have been at length in [16].

Language learning at various levels can be explored, and those that have been developed include spelling and word recognition, and the SalinLahi learning environment for Filipino heritage learners [5]. Figure 3 shows a sample screenshot of SalinLahi, showing Filipino words as well as the feedback generated by the system regarding a learner's answer.

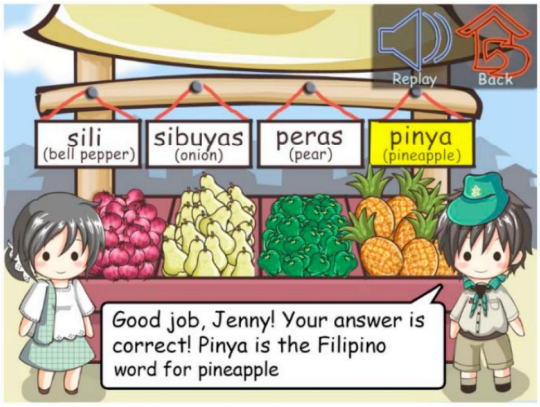

Figure 3. A sample screenshot of SalinLahi

Intelligent tutoring systems incorporate an automated tutor that embodies some characteristics of a human tutor in order to adapt its teaching style and delivery content to the comprehension level of an individual learner. For instance, Popsicle [13], a system for English second-language learners, identifies and corrects language errors committed by students while they are learning English. The tutoring response generated by the system is adjusted to the current English language proficiency level of the individual student.

MeSCh (Measurement System for Children's Reading Comprehension) measures children's reading comprehension through the automatic generation of multiple-choice questions from children's literature [12]. Questions include W-questions (who, what, when, and where), sequence questions (detailing which came first), vocabulary questions. Principles in instructional assessment are considered such as the formulation of 4W questions and the selection of distractors through the use of entries in WordNet that relate with the correct answer.

Dialogue systems can also be a kind of learning systems that develop user communication skills. For instance, *HelloPol* is a system where the user can engage the computer in a healthy dialogue in English within the political domain [3]. It is an adaptive question-answering system in that it considers in its responses the user's topic preference during the course of the dialogue.

Another emerging area in language learning is in the field of creative natural language processing where techniques in artificial intelligence and knowledge representation are utilized to produce narrative discourse in the form of stories and computational humor. For example, *Picture Books* generates stories for children from an input picture containing the background and a set of characters and object stickers [19]. A (manually-created) ontology containing relevant objects and concepts and their relations is used to generate a fable-type of story suitable for children age 4-6 years. A sample user-created picture is shown in Figure 4 with the corresponding computer-generated story at the right-hand side of the screen.

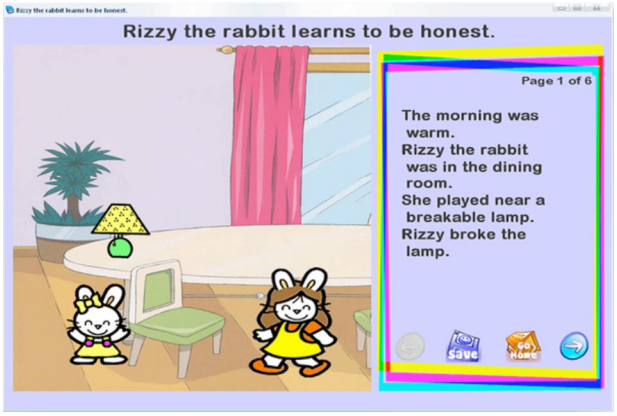

Figure 4. A sample generated story of Picture Books

A Filipino surface realizer, FilSuRe, is currently being developed to produce BookLat, a Filipino version of Picture Books that generates stories in the Filipino language. Picture Books is also being enhanced with a voice narration capability. The narrator agent will utilize an existing text-to-speech synthesizer to empathically narrate the generated story to emphasize emotions while sustaining a child's interest as well as develop his/her literacy skills.

Picture Books 2 [4] extends the scope of its predecessor by providing an environment for children to enhance their creativity through the specification of multiple scenes comprising an input picture. This Story Editor where the child defines the elements of the input picture is shown in Figure 5. The system then generates fable stories for children age 6-8 years old from the given input

picture, utilizing a manually created ontology that has now been extended with concepts on object movement and transitions between scenes. The embodiment of traits in the main character is one of the key factors that drives the flow of the story, in order to produce stories with increased character believability.

Figure 5. Story Editor of Picture Books 2

In computational humor, *Pun World* [1] makes use of *TPEG* [14] to provide a learning environment where children can enhance their spelling and vocabulary skills through wordplay. Pun World allows teachers to specify a set of training puns that are then fed to TPEG, which can automatically identify word relationships in these training puns using available linguistic resources like WordNet and ConceptNet. TPEG then uses the learned knowledge to construct its own puns which are used to teach children about pronunciation and rhyming words, spelling changes, and semantic similarities and differences between words. Figure 6 shows a sample lesson from Pun World. Table 1 shows some puns generated by TPEG.

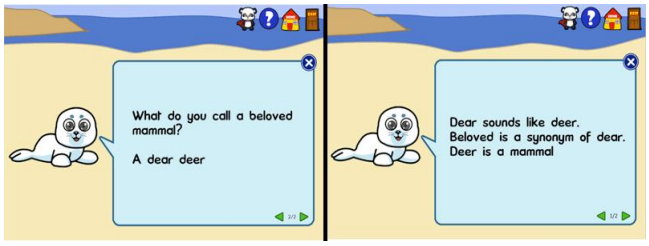

Figure 6. Sample lesson from Pun World

Table 1. Sample puns generated by TPEG

[1] What do farmers study? Transplant.
[2] What part of a man lengthens the most? The shadow.
[3] What verses are endless? The universe.
[4] What court never moves? An eyewitness.
[5] How is a camera like an eye? They both have lens.

Another application developed for composition writing is the automatic essay evaluator [8]. The system evaluates essay-type documents for mechanics, organization, and content. The grammar is checked using rule-based natural language parsing, while the latent semantic analysis technique is used to evaluate the content. The system was trained on corpora containing pre-graded essays gathered from a particular high school class, which were graded by at least two human teachers.

An emerging field known as empathic computing considers and integrates into the learning process the detection of emotions in determining the feelings of users. Emotion detection can be from non-verbal and verbal cues, and can be detected using video and/or audio files using digital signal processing [9]. For instance, while using a computer system, empathic computing provides useful information about the user's emotions; say if the user is already confused, angry, happy or sad. The concepts of empathic computing that detects non-verbal cues in sign languages such as the Filipino Sign Language (FSL) can also be employed. Currently, technologies in digital signal processing have been employed for number recognition in FSL [18].

3. LAW AND BUSINESS

An interesting NLP application is in information extraction (IE). IE involves the automatic extraction of structured data from unstructured data; that is, from long textual documents to databases. LegalTRUTHS [6] is such an IE application which aims to minimize the user's need of going through countless number and infinitely long legal documents and court decisions to extract key information about the case at hand. Figure 7 shows a sample screenshot of the relevant information extracted into table form. Overall results show precision at 91%, recall at 99%, and F-measure at 95%.

Case Number:	G.R. no. 102706
Case Title:	PEOPLE OF THE PHILIPPINES, plaintiff-appellee, vs. LEON LUMILAN, ANTONIO GARCIA and FRED ...
Division:	SECOND DIVISION
Date of Promulgation:	January 25, 2000
Ponente:	DE LEON, Jr., J.
Cases Cited	People v. Quijada, 259 SCRA 191, 232 (1996). Italics retained. Emphasis supplied.
	People vs. Parazo, 272 SCRA 512, 520-521 (1997).
	People vs. Navales, 266 SCRA 569 590 (1997).
Laws Cited:	Article 248 of the Revised Penal Code
	Article 6 of the Revised Penal Code
Decision:	Reversed
Origin of the Case:	Branch 16 of the Regional Trial Court of Isabela
Votes:	bellosillo, mendoza, quisumbing and buena, jj..
Crime(s):	attempted murder
	illegal possession of firearms
	murder
Automatic Review or Appeal:	Appeal
Penalty:	n/a
Date of Commission:	October 12, 1987
Time of Commission:	4:00 o'clock in the afternoon
Place of Commission:	in the Municipality of Ilagan, Province of Isabela

Figure 7. LegalTruths sample IE output

In governance, the eParticipation or eLegislation [17] is an ongoing project that integrates NLP applications towards greater constituent participation in coming up with legislation. It has an IE system and an opinion clustering system. The IE system allows the users to be able to work with structured concise data from the documents of Philippine Senate or Congress. While allowing these documents and the corresponding data to be viewed by the public in general and, specifically, the elected officials' constituents, the elected officials can consequently

gather their constituents' comments on particular proposed laws as these are being drafted, through online forums. These comments can accordingly be clustered and grouped together in a systematic and automatic way so that the elected officials at any point in time can have a grasp of how people feel and what they think about these proposed laws.

In business, NLP can find applications as a question-answering system to allow users to specify questions in natural language and receive corresponding answers also in their own language. QA4BI [7] is one such system that interprets and answers comparative and evaluative questions under the domain of business intelligence. The system is currently able to evaluate at least 15 predicates from data gather manually from online biotechnology news site (BioSpace.com).

4. SUMMARY

In this paper, I have presented some of the practical applications of human language technology (HLT) researches that we have conducted in the country. The discussions show that although the HLT field is a highly technical field, it also runs across various other practical fields of endeavors in other fields such as in culture and the arts, education, law and business. The development of these applications for HLT in the country provides a very rich and diverse platform for further research and study, opening up many emerging areas and fields of study.

5. REFERENCES

[1] Aban, V.R., Fernandez, T.J., Pascual, M.C., & Ong, E. (2010). Exploring Puns for Spelling and Vocabulary Enrichment. In *Proceedings of the 7th National Natural Language Processing Research Symposium*, 27-32, November 19-20 2010, De La Salle University, Manila.

[2] Acosta, J. I., Espiritu, R. V., Ngo, C. J., & Wong, J. (2008). *SpeL-IT: House on Phonic Hill*. Unpublished undergraduate thesis, De La Salle University, Manila, the Philippines.

[3] Alimario, P. M., Cabrera A., Ching, E., Sia, E. J., & Tan, M. W. (2003). HelloPol: An adaptive political conversationalist. In *Proceedings of the 1st National Natural Language Processing Research Symposium*. Manila, the Philippines: De La Salle University, College of Computer Studies.

[4] Ang, K., Antonio, J., Sanchez, D., Yu, S., & Ong, E. (2010). Generating Stories for a Multi-Scene Input Picture. In *Proceedings of the 7th National Natural Language Processing Research Symposium*, 21-26, November 19-20 2010, De La Salle University, Manila.

[5] Cheng, C. K., Antay, M. R., Jumarang, M., Regalado, R. V., & Fernandez, L. M. (2010). *SalinLahi*: A Web-Based, Interactive Learning Environment for the Filipino Language. In *Proceedings of the First International Conference on Heritage/Community Languages*. Los Angeles.

[6] Cheng, T., Cua, J., Tan, M. D., Yao, K. G., & Roxas, R. (2009). *Information Extraction from Legal Documents*, Proceedings of the Symposium on NLP 2009 (IEEE), Bangkok, Thailand, October 21-23, 2009.

[7] Choi, K., Pacana, R. M., Tan, A. L., Yiu, J., & Lim, N.R. (2010). A Question Answering System that Performs Evaluations and Comparisons on Structured Data for Business Intelligence in Biotechnology. In *Proceedings of the 7th National Natural Language Processing Research Symposium*, November 19-20, De La Salle University, Manila.

[8] Cruz, M, Escutin, M., Estioko, A., & Plaza, M. (2003). Automated Essay Evaluator. In *Proceedings of the 1st Natural Language Processing Research Symposium*. Manila, the Philippines.

[9] Cu, J. & Roxas, R. (2009). *Speech Corpora and Applications: Philippine Country Report*, Proceedings of the Oriental-COCOSDA Conference, August 10-11, 2009.

[10] Dimalen, D. & Roxas, R. (2007). *AutoCor: A Query-based Automatic Acquisition of Corpora of Closely-Related Languages.* 21st Pacific Asia Conference on Language, Information and Computation (PACLIC-21), Seoul, Korea, November 1-3, 2007. A Full Paper Presentation. Index to Scientific & Technical Proceedings (ISTP).

[11] Dita, S., Roxas, R., & Inventado, P. (2009). Building online corpora of Philippine languages. In *23rd Pacific Asia Conference on Language, Information and Computation* (PACLIC-23) (pp. 646-653). Hong Kong SAR, China.

[12] Fajardo, K., Di, S., Novenario, K., & Yu, C. (2008). *Mesch: Measurement system for children's reading comprehension.* Unpublished undergraduate thesis, De La Salle University, Manila, the Philippines.

[13] Gurrea, A. M., Liu, A., Ngo, D., Que, J., & Ong, E. (2006). Recognizing Syntactic Errors in Written Philippine English. In *Proceedings of the 3rd National Natural Language Processing Research Symposium*. Manila, the Philippines.

[14] Hong, B.A., & Ong, E. (2009). Automatically Extracting Word Relationships as Templates for Pun Generation. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity*, 24-31, June 5 2009, Boulder, Colorado, USA.

[15] Roxas, R. (2007). *e-Wika: Philippine Connectivity through Languages.* Lecture Presented at the 4th National Natural language Processing Symposium, De La Salle University, June 14-16, 2007. pp 12-18.

[16] Roxas, R., Alcantara, D., & Borlongan, A. (2010). *Language Documentation and Applications in the Philippines: Implications for Mother Tongue-Based Multilingual Education*, Philippine Education Research Journal, 2010.

[17] Roxas, R., Borra, A., Cheng, C., & Ona, S. (2010). eParticipation http://panegov.net/projects/project1.htm

[18] Sandjaja, I., & Marcos, N. (2009). Sign language number recognition. In *Proceedings of the NCM 2009 (International Conference on Networked Computing, Advanced Information Management and Digital Content and Multimedia Technologies), 5th International Joint Conference on INC, IMS and IDC,* August 25-27, 2009. Seoul, Korea.

[19] Solis, C., Siy, J. T., Tabirao, E., & Ong, E. (2009). Planning author and character goals for story generation. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity*, 63-70, June 5 2009, Boulder, Colorado, USA.