Computational Linguistics Research on Philippine Languages

Rachel Edita O. ROXAS
Software Technology Department
De La Salle University
2401 Taft Avenue, Manila, Philippines
ccsror@ccs.dlsu.edu.ph

Allan BORRA
Software Technology Department
De La Salle University
2401 Taft Avenue, Manila, Philippines
ccsabb@ccs.dlsu.edu.ph

Abstract

This is a paper that describes computational linguistic activities on Philippines languages. The Philippines is an archipelago with vast numbers of islands and numerous languages. The tasks of understanding, representing and implementing these languages require enormous work. An extensive amount of work has been done on understanding at least some of the major Philippine languages, but little has been done on the computational aspect. Majority of the latter has been on the purpose of machine translation.

1 Philippine Languages

Within the 7,200 islands of the Philippine archipelago, there are about one hundred and one (101) languages that are spoken. This is according to the nationwide 1995 census conducted by the National Statistics Office of the Philippine Government (NSO, 1997). The languages that are spoken by at least one percent of the total household population include Tagalog, Cebuano, Ilocano, Hiligaynon, Bikol, Waray, Pampanggo or Kapangpangan, Boholano, Pangasinan or Panggalatok, Maranao, Maguin-danao, and Tausug.

Aside from these major languages, there are other Philippine dialects, which are variants of these major languages. Fortunato (1993) classified these dialects into the top nine major languages as above (except for Boholano which is similar to Cebuano).

2 Language Representations

Linguistics information on Philippine languages are extensive on the languages mentioned above, except for Maranao, Maguin-danao, and Tausug, which are some of the languages spoken in Southern Philippines. But as of yet, extensive research has already been done on theoretical linguistics and little is known for computational linguistics. In fact, the computational linguistics researches on Philippine languages are mainly focused on Tagalog.[1] There are also notable work done on Ilocano.

Kroeger (1993) showed the importance of the grammatical relations in Tagalog, such as subject and object relations, and the insufficiency of a surface phrase structure paradigm to represent these relations. This issue was further discussed in the LFG98, which is on the problem of voice and grammatical functions in Western Austronesian Languages. Musgrave (1998) introduced the problem certain verbs in these languages that can head more than one transitive clause type. Foley (1998) and Kroeger (1998), in particular, discussed about long debated issues such as nouns in Tagalog that can be verbed, the voice system of Tagalog, and Tagalog as a symmetrical voice system. Latrouite (2000) argued that a level of semantic representation is still necessary to explicitly capture a word's meaning.

Crawford (1999) contributed to an issue on interrogative sentences and suggested that the restriction on wh-movement reveals the syntactic structure of Tagalog.

Potet (1995) and Trost (2000) provided general materials on computational morphology, though, both presented examples on Tagalog.

Rubino (1997, 1996) provided an in-depth analysis of Ilocano. Among the major contributions of the work include an extensive treatment of the complex morphology in the language, a thorough treatment of the discourse

[1] Tagalog (or Pilipino) has the most number of speakers in the country. This may be due to the fact that it was officially declared the national language of the Philippines in 1946.

particles, and the reference grammar of the language.

3 Applications in Machine Translation

Currently, most of the empirical endeavours in computational linguistics are in machine translation.

3.1 Filipino MT Software

There are several commercially available translation software, which include Philippine language, but translation is done word-for-word. One such software is the Universal Translator 2000, which includes Tagalog among 40 other languages. Although omni-directional, translation involving Tagalog excludes morphological and syntactic aspects of the language

Another software is the Filipino Language Software, which includes Tagalog, Visayan, Cebuano, and Ilocano languages.

3.2 Machine Translation Research

IsaWika! is an English to Filipino machine translator that uses the augmented transition network as its computational architecture (Roxas, 1999). It translates simple and compound declarative statements as well as imperative English statements. To date, it is the most serious research undertaking in machine translation in the Philippines.

Borra (1999) presented another translation software that translates simple declarative and imperative statements from English to Filipino. The computational architecture of the system is based on LFG, which differs from IsaWika's ATN implementation. Part of the research was describing a possible set of semantic information on every grammar category to establish a semantically-close translation.

4 Conclusion

There are various theoretical linguistic studies on Philippine languages, but computational linguistics research is currently limited. CL activities in the Philippines had yet to gain acceptance from its computing science community.

References

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

Crawford, C (1999) *A Condition on Wh-Extraction and What it Reveals about the Syntactic Structure of Tagalog*. http://www.people.cornell.edu/pages/cjc26/l304final.html

Foley, B (1998) *Symmetric Voice Systems and Precategoriality in Philippine Languages*. In LFG98 Conference, Workshop on Voice and Grammatical Functions in Austronesian Languages.

Fortunato, Teresita, *Mga Pangunahing Etnoling-guistikong Grupo sa Pilipinas*, 1993.

Kroeger, P (1998) *Nouns and Verbs in Tagalog: A Response to Foley*. In LFG98 Conference.

_____ (1993) *Phrase Structure and Grammatical Relations in Tagalog.* CLSI Publications, Center for the Study of Language and Information, Stanford, California.

Latrouite, Anja (2000) *Argument Marking in Tagalog*. In Austronesian Formal Linguistics Association 7th Annual Meeting (AFLA7). Vriji Universiteit, Amsterdam, The Netherlands.

Musgrave, S (1998) *The Problem of Voice and Grammatical Functions in Western Austronesian Languages*. In LFG98 Conference.

National Statistics Office (1997) "*Report No. 2: Socio-Economic and Demographic Characteristic"*, Sta Mesa, Manila.

Potet, J (1995) *Tagalog Monosyllabic Roots*. In Oceanic Linguistics, Vol. 34, no. 2, pp. 345-374.

Roxas, R., Sanchez, W. & Buenaventura, M (1999) *Final Report of Machine Translation from English to Filipino: Second Phase*. DOST/UPLB.

Rubino, C (1997) *A Reference Grammar of Ilocano*. UCSB Dissertation, UMI Microfilms.

_____ (1996) *Morphological Integrity in Ilocano*. Studies in Language, vol. 20, no. 3, pp. 333-366.

Trost, Harald (2000) *Computational Morphology*. http://www.ai.univie.ac.at/~harald/handbook.html