Introducing Filipino Wordnet

Allan Borra
Center for Language Technologies
College of Computer Studies
De La Salle University, Manila, Philippines
borgz.borra@delasalle.ph

Adam Pease
Articulate Software
Angwin, CA, USA
apease@articulatesoftware.com

Rachel Edita. O. Roxas
Center for Language Technologies
College of Computer Studies
De La Salle University, Manila, Philippines
rachel.roxas@delasalle.ph

Shirley Dita
Department of English and Applied Linguistics
College of Education
De La Salle University, Manila Philippines
shirley.dita@dlsu.edu.ph

Abstract

In this paper, we introduce the Filipino wordnet project (FilWordNet). Filipino is the national language of the Philippines spoken by some 90 million people as their first or second language. However, it has historically had a limited number of computational linguistics resources. Creating the Filipino wordnet can be seen as the first step to enable a wide range of research projects. We describe our process of building a wordnet, including issues with the Filipino language itself, its morphology and structure.

1 Introduction

We discuss the construction of a WordNet for Filipino. Morphology is discussed to establish the need for analyzers and generators to support root word entries in the Wordnet as well as sysnset entries in root word form. Other aspects are investigated such as idiosyncratic and culturally unique words in Filipino.

2 Background

The motivation for the creation of Filipino Wordnet is to provide a solid base of formal linguistic information that could subsequently be used for pertinent language technology applications as outlined by (Morato et al., 2004) . These include information retrieval and extraction, particularly in concept identification in natural language and in query expansion, language teaching, translation applications, and in parameterizable information systems which allowed personal searching of documents based on users' interests. While there has been at least one earlier proposal for a Filipino WordNet (Tan&Lin, 2007), the proposed work was not performed.

2.1 Filipino Language

Tagalog is a language in the Austronesian group of languages, and is, de facto, the basis for the national language of the Philippines, called Filipino. Tagalog is a free word order language, and is somewhat agglutinative with a rich set of affixes, and as such, so is Filipino. The interested reader is referred to (Schacter&Otanes, 1972) for more information about Tagalog grammar..

As in many wordnet projects, the first step is to determine what words deserve to be synsets and which are morphological extensions of the root.

To take cases in English to illustrate this, 'walked' is simply the past tense of 'walk', and does not have additional meaning. In contrast, while 'catfish' is a fish that looks somewhat like a cat, there is no automatic way to know that is the case, and not that it is a cat that looks like a fish, or a cat that likes fish. These facts would encourage an English wordnet to include 'walk' but not 'walked', and to include 'cat', 'fish' and 'catfish' all as separate synsets.

The following examples illustrate morphological phenomena in Filipino. In each case, we illustrate different affixes that affect the focus of the verb, as well as enumerating the different tenses or aspect for each verb with a given focus. Verb focus has no direct analog in English, other than possibly a prosodic emphasis (Szwedek, 1986). Verb focus indicates to the listener where to place the focus of attention in a sentence

Take for example the case of the root word *bili* 'to buy' in actor focus. This focus indicates to the listener that the attention should be on the performer of the action. For actor focus, Filipino uses the infix –um-

*B**um**ili* = 'bought' – perfective (*Bumili ang bata ng kendi.* 'The child bought a candy.')

*B**um**ibili* = 'is/are buying' – the -um- focus marker + consonant-vowel reduplication yield progressive aspect (*Bumibili ang bata ng kendi*. 'The child is buying a candy.').

Bibili = 'will buy' – consonant-vowel

reduplication yields imperfective (or contemplated, unrealized) aspect (*Bibili ang bata ng kendi*. 'The child will buy a candy.') Note that rather confusingly for the non-native speaker, the infix disappears in this aspect.

In each case, one can imagine an English speaker placing an emphasis through loudness or pitch on 'child'.

For benefactive focus Filipino uses a prefix i- + infix -in-. Note that the future tense shifts the affix to the end of the word. Benefactive focus indicates that the focus of attention of the listener should be on the entity that benefits from the action.

Ibin*ili* = 'bought' - perfective (*Ibinili ng bata ng kendi ang beybi.* 'The child bought a candy for the baby.')

Ibin*ibili* = 'is/are buying' – progressive/ on-going – (*Ibinibili ng bata ng kendi ang beybi*. 'The child is buying a candy for the baby.')

I*bibili* = 'will buy' – imperfective – (*Ibibili ng bata ng kendi ang beybi*. 'The child will buy candy for the baby.')

In each case, one can imagine an English speaker placing an emphasis through loudness or pitch on 'baby'.

From here, we can complicate the morphology a bit by adding other morphemes, with the example *maipabibili* (I will be able to have him buy (something))

ma- = abilitative prefix (to be able to)
i- = benefactive topic marker (beneficiary of the action is the focus)
pa- = causative marker
bi = aspect marker (imperfective) - consonant-vowel reduplication form
bili = root

These examples provide an insight as to the effect of different affixes when applied to a particular root. We believe that a morphological analyzer is a better approach in modeling Filipino words than storing all of the inflections of a root word in the wordnet as different synsets.

Our initial approach is to be strict in only allowing root forms in the wordnet, unless the word has gained some meaning that cannot be automatically deduced from the root and any affixes.

Uniquely Filipino words

It is very often the case that each new wordnet will have synsets that do not appear in most or even any other existing wordnet (Elkateb et al, 2007). This is also true with Filipino. Let us take a few examples.

tinikling – a cultural dance that originated in the Visayas region utilizing two moving bamboo sticks over which the dancers perform
bayanihan – the spirit of communal unity
bilas – spouse of the brother or sister of one's own spouse
hilamos – to wash one's face

Words such as these form part of the motivation for using a formal ontology. While some wordnets have used English as an interlingua and created phrases to stand in the place of otherwise unlexicalized concepts, in our work, we use SUMO as an interlingua which can contain concepts which stand for the lexicalized concepts of any particular language. For example, rather than add a new English synset corresponding to "spouse of the brother or sister of one's own spouse", we create a concept in SUMO with that definition and relate the Filipino synset to it. This avoid creating synsets in a given language that are "artificial" and not actually lexicalized units. To use the example above of "hilamos", consider Figure 1

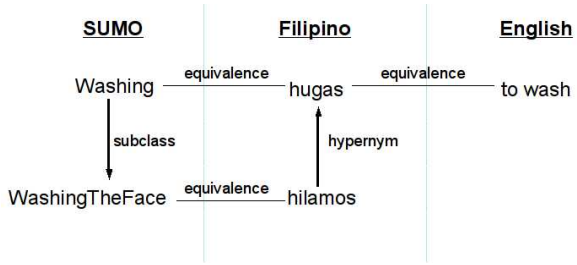

Figure 1: Relation among FilWordNet, SUMO and Princeton's English WordNet

"Hilamos" is a word not lexicalized in English, so above we show it as a term only linked to SUMO.

We should note that the equivalence relation is an informal one. It is neither wordnet semantic link nor a formal logical link as would be found in SUMO, rather it is a relationship without strict definition but denoting intuitively very close

similarity to the point of equivalence. We also show the relationship subclass in SUMO which is a formal and truth-preserving relationship, and the hypernym relationship which is one of wordnet's semantic links.

3 Wordnets

Since Princeton's WordNet (PWN) is well-known, it may be sufficient simply to refer the reader to (Fellbaum, 1998). For the purposes of this paper, it bears mentioning that there are several features of WordNet that make it an ideal model for our work with FilWordNet, and an important product to link to.

- PWN is a mature product, having been started over two decades ago (Miller, 1985)
- It is very comprehensive, with over 115,000 word senses, making it the largest wordnet in existence
- It has been free since the project's inception
- It is richly interconnected as a semantic network
- Many other languages have linked their wordnet projects to it manually

4 Suggested Upper Merged Ontology

The FilWordNet project will provide a deep semantic underpinning for each psycholinguistic concept. We take the same approach that was previously used in mapping all of PWN to a formal ontology (Niles & Pease, 2003), the Suggested Upper Merged Ontology (Niles & Pease, 2001), as well as more recently using SUMO as the formal underpinning for Arabic WordNet (Elkateb et al 2007)

Synsets map to a general SUMO term or a term that is directly equivalent to the given synset (Figure 1). New formal terms will be defined to cover a greater number of equivalence mappings, and the definitions of the new terms will in turn depend upon existing fundamental concepts in SUMO. The process of formalizing definitions will generate feedback as to whether word senses in WN need to be divided or combined and how the glosses may be clarified. Since many wordnets in other languages are already linked by synset number, this work will benefit wordnets in other languages as well.

The Suggested Upper Merged Ontology (SUMO) (Pease&Niles, 2002),(Niles&Pease, 2001) is a freely available, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format (SUO-KIF) (Pease, 2000), and also has a necessarily lossy translation into the OWL semantic web language. It has undergone nine years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. SUMO has been extended with a number of domain ontologies, which are also public, that together number some 20,000 terms and 70,000 axioms. SUMO has been mapped by hand to the WN lexicon of over 115,000 noun, verb, adjective and adverb senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list. SUMO is the only formal ontology that has been mapped to all of WN, and the only formal upper ontology that has been extended with a number of domain ontologies that are also open source. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUMO-KIF and SUMO to be expressed in multiple natural languages. These include English, German, Arabic, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin).

The ontology as a structured ILI

The comprehensive mapping and definition of synsets in FilWordNet to SUMO concepts reinforces a new perspective on the role of an Interlingual Index (ILI) in connecting wordnets (Vossen, 2004, Vossen et al 1999, Vossen 1998).

In the FilWordNet project, we want to take this idea a step further, as was done with Arabic. If both FilWordNet and English WN synsets are exhaustively defined in terms of SUMO concepts, SUMO can in effect become the ILI for wordnets. This means that SUMO not only maps word meanings and synonyms across languages

but also provides a formal semantic framework for all these languages.

The development of FilWordNet will include a transition phase where FilWordNet synsets are both linked to the English WN serving as an ILI and exhaustively defined with SUMO.

5 Project Description

FilWordNet began an introductory set of lectures to students and faculty at De La Salle University, Manila, on wordnets, linguistics semantics and formal semantics. Six students volunteered to be the actual creators of the synsets with an intensive two-week process to complete the project. The objective was to create 40 synsets a day per person with each student present for roughly three hours a day. We were able to achieve an initial set of 1,000 synsets, although full completion took a few weeks longer than anticipated due to external unrelated events. The students are expected to continue work in order in a few months to cover the approximately 4,600 base concepts (Pease et al, 2008). Additionally, a small cash prize was announced for the creators of the most synsets in hopes of creating a mature wordnet of greater than 10,000 synsets at the end of the academic year 2009-2010.

We created an initial seed list of Princeton's English WordNet synsets for students to get started. This consisted of some synsets chosen from intuition as being semantically distant from one another. Each student was expected to translate synset word names and definitions into Filipino. They were assisted in this task by the use of Calderon's Tagalog-English-Spanish dictionary (Calderon, 1915). After some collaborative work on this initial set, they expanded their set to hypernyms and hyponyms of the seed word and continued in this fashion.

We have used Princeton WordNet semantic links and assumed them to be correct in Filipino subject to manual verification later on. Similarly, we reuse the links from Princeton WordNet to SUMO also subject to later manual verification as to whether it is also valid for Filipino.

We treat the links to SUMO and semantic links within wordnet to be an important part of the quality assurance process for wordnet construction. Considering the semantic links between different synsets helps the lexicographer to determine whether the definition and synset grouping is valid. For example, a critical test is to look amongst sibling synsets and consider whether the definition of a given synset fully distinguishes it from its siblings.

Initially, we had the students create their translations simply in spreadsheets tracking the link to English via each synsets WordNet 3.0 synset number. We have installed DebVisDic (Horak et al, 2006) and will be migrating to that tool as our construction environment shortly. We expect that this will help considerably especially with respect to group coordination.

We plan an open source release of FilWordNet for early in 2010, once we are close to covering the base concepts.

6 Conclusions and Future Work

FilWordNet is an enabling resource for computational linguistics on Filipino. We are currently conducting linguistics research on the evolution of Tagalog grammar among metropolitan residents of the Philippines in which we plan to use FilWordNet in performing manual markup of Filipino corpora. FilWordNet will be a basis for a stemmer/lemmatizer that will use FilWordNet to prevent overly "greedy" removal of affixes from words. FilWordNet will also provide a basis for work in developing a named entity recognition system. With a series of projects that all leverage the work on FilWordNet, we hope that will create motivation to continue expanding and improving this product. Additionally, we hope to involve other universities in the Philippines in this effort to improve linguistic resources for the national language.

Acknowledgments

We would like to thank our students, Alvin Garcia, Hun Ping Yu, Bryan Lacaden, Jeremy Bondoc, Darren So and Jhovee Yap who have created the actual synsets. We are grateful for their efforts. We acknowledge De La Salle University, Manila for hosting Adam Pease as a visiting scholar, and the Philippine Council for Advanced Science and Technology Research and Development (PCASTRD) of the Department of Science and Technology (Philippines) and Commission on Higher Education (Philippines) for partial funding of his visit.

References

Calderón, S., (1915) Diccionario Ingles-Español-Tagalog, Con partes de la oracion y pronunciacion figurada. Primera Edición, Manila, Libreria y Papeleria de J. Martinez, Plaza P. Moraga 34/36, Plaza Calderón 108 y Real 153/155, Intramuros. See also http://www.gutenberg.org/etext/20738

Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Building a WordNet for Arabic, in Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006).

Fellbaum, C., (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Horak, A., Pala, K., Rambousek, A., and Povolny, M. (2006) DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool. In Proceedings of the Third International WordNet Conference - GWC 2006. Brno, Czech Republic: Masaryk University, pp. 325-328. ISBN 80-210-3915-9.

Miller, G., (1985) "WordNet: a dictionary browser." In Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo.

Morato, J., Marzal, M.A., Llorens, J., & Moreiro, J (2004). WordNet Applications. In Proceedings of the Second Global WordNet Conference (GWC-2004). Brno, Czech Republic.

Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.

Niles, I., and Pease, A. (2001). Towards a Standard Upper Ontology. In: Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9. See also http://www.ontologyportal.org

Pease, A., (2000). Standard Upper Ontology Knowledge Interchange Format. Web document http://suo.ieee.org/suo-kif.html. This is largely a condensed version of the language described in (Genesereth, 1991)

Pease, A., (2003). The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.

Pease, A., Fellbaum, C., and Vossen, P., (2008) Building the Global WordNet Grid. Proceedings of the CIL-18 Workshop on Linguistic Studies of Ontology, Seoul, South Korea.

Schachter, P., and Otanes, F., (1972) Tagalog reference grammar. ISBN: 0520017765, Berkeley : University of California Press

Szwedek, A (1986). A linguistic analysis of sentence stress, Gunter Narr Verlag: Tubingen , ISBN: 3-87808-298-3

Tan, P., and Lim, N., (2007) FilWordNet: Towards a Filipino WordNet. 4th National Natural Language Processing Research Symposium Proceedings, CSB Hotel, June 14-16, 2007, ISSN 1908-3092

Vossen, P. (ed) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.

Vossen, P. Peters, W., J. Gonzalo. (1999). 'Towards a Universal Index of Meaning'. Proceedings of the ACL-99 Siglex workshop, University of Maryland, 81-90

Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. International Journal of Lexicography, Vol. 17 No. 2, OUP, pp 161-173