Towards a Sinhala Wordnet

Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe
Language Technology Research Laboratory,
University of Colombo School of Computing, Sri Lanka
{wvw, dlh, cml, ugn, arw}@ucsc.lk

Tissa Jayawardana
Department of Linguistics,
University of Kelaniya,
Sri Lanka

Abstract

This paper describes the methods adopted and the issues addressed in building a Sinhala Wordnet, based on the Princeton English WordNet (PWN). Its aim is to develop the most important parts of a wordnet for the Sinhala language, in order for it to be of optimal use without being complete. The importance of entries were estimated using the word frequencies of the 10 million word UCSC Sinhala corpus of Contemporary Sinhala, and the relevant lexico-semantic relations extracted from the PWN. The paper describes how the Sinhala Wordnet was developed with a view to presenting a recommended strategy for other languages for which wordnets may be developed in the future.

1 Introduction

Wordnet is one of the most useful lexical resources for many key natural language processing and computational linguistic tasks including Word Sense Disambiguation, Information Retrieval and Extraction, Machine Translation, and Question Answering among others. It is based on the theories developed in Lexical Semantics and defines different senses associated with the meaning of a word and other well-defined lexical relations such as synonym, antonym, hypernym, hyponym, meronym and holonym.

The Princeton WordNet (PWN) (Fellbaum, 1998), is a large lexical resource developed for English, which contains open class words namely; nouns, verbs, adjectives and adverbs. These words have been grouped together based on their meanings, with a single set of such synonyms being called a *synset*. Many efforts have been reported in recent years to develop such lexical resources for other languages (e.g. Darma Putra et. al. (2010), Elkateb at al, (2006) among others) based on the relations defined in the PWN.

Sinhala is an Indo-Aryan language spoken by a majority of Sri Lankans. It is also one of the official and national languages of Sri Lanka. The University of Colombo School of Computing (UCSC) has been involved in building Sinhala language resources for NLP applications for many years. Some of these include a 10 million word Sinhala corpus, a part-of-speech tag set, and a tri-lingual dictionary. The motivation behind the project to build a Sinhala wordnet is to fulfill the requirement of a semantico-lexical resource for NLP applications.

A brief overview of three prominent wordnet projects namely the PWN, the *Euro WordNet* (Vossen, 2002), and the *Hindi WordNet* (Narayan et. al. (2002) and Chakrabarti and Bhattacharyya (2004)) were closely examined as a part of the Sinhala wordnet development project, to understand the approaches taken, structures used, language specific issues and the functionalities available in them. Using this input, it was decided to define the relations among Sinhala words using PWN sense IDs in order to keep the consistency with many other wordnet initiatives in the interest of possible interoperability. This also helped in developing the Sinhala wordnet with less effort, by using the linguistic notions that held across languages and language families.

The initial idea that the PWN synsets could be directly translated for use as the Sinhala Wordnet had to be abandoned owing to the top-level categories in it being less relevant in tasks such as word-sense disambiguation owing to the lack of

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

ambiguity in them in the Sinhala language. Instead, the UCSC Sinhala corpus, which consists of 10 million Sinhala words in contemporary use, was used as the main resource to base the selection of the most important parts of the wordnet which needs to be built to be of use for applications for Sinhala. High frequency open class words from the corpus were identified in order to discover word senses that contributed most to contemporary language use. Each of these words was then considered as a candidate for inclusion in the Sinhala wordnet sub-set to be constructed first. Other senses relating to these words were then enumerated in consultation with language and linguistics scholars. This strategy helped to build the Sinhala wordnet in a phased manner, starting with most prominent and hence multi-sense words in the language.

This paper presents the work carried to develop the Sinhala wordnet using the PWN synset IDs. The rest of paper will describe the methodology, challenges and the future work of the Sinhala wordnet project.

2 Methodology

A survey of potential resources for the Sinhala wordnet project was carried out at the beginning of the project. As a result of this survey, it was found that the tradition of thesaurus building is not new to Sinhala language studies but has been in general fairly well established in traditional linguistic studies originating from ancient India.

Though there are some Sinhala language resources available in the Sinhala literature which are closer to the current work, many of these could not be directly used due to poor coverage of contemporary Sinhala (mainly covers traditional ancient language) and the poverty of concept classification (confined to religious and preliminary concepts). Having examined them thoroughly one main resource and a couple of supplementary resources were identified as primary sources for the project. A few popular Sinhala dictionaries and thesauri were among these (e.g. Wijayathunga, 2003).

The literature concerning the semantic aspect of the Sinhala language is relatively limited due to it not being handled formally by scholars of Sinhala language research. This has led to a situation where it is difficult to express the semantics of words and their sense relations accurately. In order to address these issues, it was decided to complement the information given in such Sinhala language resources in an informal manner by working with linguistic scholars who have a strong theoretical background in both traditional grammar and modern linguistic analysis of Sinhala and English languages.

Having closely studied the approaches taken in other wordnet initiatives, a strategy for the development of the Sinhala wordnet was established. Many wordnet initiatives have used a top-down approach, in which abstract concepts have been enumerated starting with a kind of upper level ontology and then gradually working down over many decades. Owing to time and resource limitations, we had to use a more data-driven approach to clearly identify the most important subset of senses within a wordnet that would be of most value to researchers. As a significant quantum of work has been done in the PWN in terms of building the infrastructure for all later wordnets, our strategy was developed in such a way that lessons learnt from the PWN project could be used to avoid most of the hurdles that have been negotiated by the developers of the PWN.

Figure 1 shows the workflow of the development process of the Sinhala Wordnet. The steps of the methodology of the Sinhala Wordnet project can be divided into sub tasks as described below.

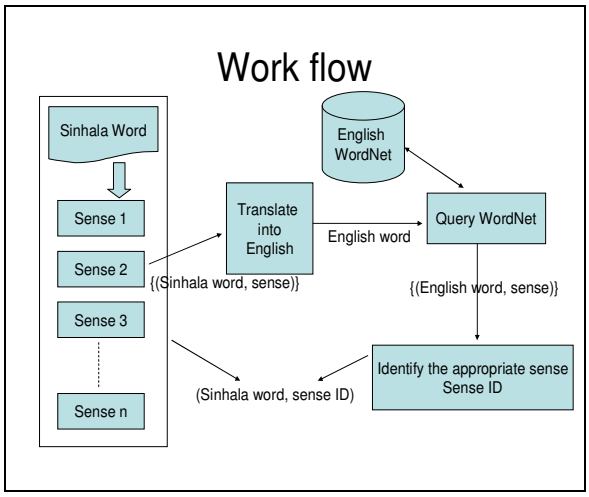

Figure 1. Workflow of Sinhala Wordnet Development

2.1 Word Selection Process

At the outset, the words to be considered for inclusion in the Wordnet were chosen from the UCSC Sinhala Corpus according to their frequency. The most frequently occurring 500 words excluding function words were chosen to

build the prototype of the Sinhala Wordnet. Next this was expanded to include the top 1000 words once the strategy was well established. As Sinhala is a morphologically rich language there are many different word forms for a given base form and only one single form called *lemma* is selected for the current system. In cases where a word form different to the base form had a different semantic value, that form was considered as a separate entry. Some words have alternate spellings and phonological variations that have led to semantic variations and such words are also considered as separate entries in wordnet.

2.2 Sense Identification Process

As discussed in Section 2.1, one word can have more than one sense and it is extremely difficult to identify all the senses of a given word. We followed two approaches to identify the senses of words, namely dictionary look up and look up of English translations of the corresponding word in the PWN. Finally, a linguistic scholar determined the list of senses for a given word after reviewing the potential senses given in the dictionary and the PWN. The main source for extracting Sinhala word senses was *Maha Sinahla Sabdakoshaya* (Wijayathunga, 2003), which is the major Dictionary of the contemporary Sinhala language.

2.3 Sense Relation Extraction

PWN defines six main word sense relations, namely, synonyms, antonyms, hypernyms, hyponyms, meronyms and holonyms. As defining them from scratch is time consuming and requires a sophisticated expertise in lexical semantics, it was decided to extract them from the PWN database and store them in a human readable format. The main motivation behind this decision was the fact that a majority of the senses are language and culture independent. Therefore this approach helps incorporating Sinhala words with relations given for English, in order to build the Sinhala wordnet with less effort.

2.4 Sinhala to English Translation and PWN Query

The accurate English translation for a given sense of a Sinhala word was determined by a linguistic scholar conversant in both Sinhala and English language usage. Having precisely translated the Sinhala word sense into English, it is in turn looked up in the PWN to obtain the relevant *synset identifier*.

2.5 Sense ID Assignment

The Sinhala word with a particular word sense is then inserted to the Sinhala wordnet database with the sense identifier obtained according to the step described in 2.4. This process helped to maintain all the sense relations, which have already been defined in the PWN database, automatically and with no extra effort on our part.

2.6 Gloss Translation

After identifying the exact sense ID for a given word-sense, we used the knowledge of expert translators to translate the gloss defined in the relevant PWN entry into Sinhala. Translators were given the freedom to change the gloss according to the language-culture of Sri Lankan Sinhala, when the PWN gloss was found to be not appropriate for the context.

2.7 Synset Identification

When the sense ID, POS and the gloss was determined for a given sense, native speakers knowledge and the other resources such as dictionaries and thesauri were used to identify the corresponding synset for that sense. This was manually done by two language experts.

The senses indentified through above process were stored in an Excel sheet (Figure 2) and currently has not been integrated with any user interface. More details on data storage are explained under Future Work.

Synset	PWN ID	POS	Gloss
සඳහනාසවහනාවිස්තරය	06418196	n	කෙටි ලිවීන සටහන
ඉන්නවා	02703136	v	තැනක හෝ තත්වයක නැවතී සිටීනවා
මූලික	01922424	s	මූලික දත්ත හෝ සුනිපත්තිවලට අදාළ වන්නා වූ
ජීවිතය	13777175	n	ජීවත්වීමේ ජෛසර්ගික ආකාරය හෝ රටාව
වැඩ;රාජකාරිය;කාර්යය;කටයුත්	00570312	n	යමක් කිරීම සඳහා හෝ සැදීම කෙරේ යොමු වූ ක්‍රිය
ගෙදර	08373013	n	ඔබ ජීවත් වන රට, ප්‍රාන්තය හෝ නගරය
අදහස;සිතුවිල්ල;චින්තනය	05761049	n	සංකල්පනාව, ප්‍රජානනයේ අන්තර්ගතය
සමණ;කරම;ගණනාරාශිය;වපස	00032028	n	සුමාණ කළහැකි යමකින් ඇති සුමාණය
සූරනවා	00479055	v	අවසාන කරනවා හෝ අවසානයට පැමිණෙනවා
චිනාශය	07233906	n	යමක් සම්පූර්ණයෙන් සිඳ බිඳ දමන සිද්ධිය (හෝ සි
මහාභැතිම;අදහස	05841869	n	යම් විශ්වාසයක් තබනු ලැබූ අනුමාන අදහසක්
මූලික	01051890	s	ආරම්භයක් ලෙස හෝ පදනමක් ලෙස සේවය කරු
නිර්මාණය කරනවා	01607166	v	කෘතිම නිෂ්පාදනයක් ඇති කරනවා
සාමාන්‍යයෙන්	00491749	r	ස්වභාවික හෝ සාමාන්‍ය අන්දමට
නැවත;ආපසු;ආයිත්;ආයිමත්;ආ	00041086	r	ආයෙමත්, අස්තින්
ජාතිකාදෑ	03070805	a	ජාතියකට හෝ රටකට අයත්
දෙනා;විස්සිඋළුදෙන	01868513	n	ඔධනුව, විශේෂයෙන් ක්ෂීරපායී සතුන්ගේ ස්තී ලිංග

Figure 2. Sinhala Wordnet Database

3 Challenges for 'new' languages

Several linguistic issues need to be addressed in order to capture language specific features in the design of the system. Most of these occurred owing to the morphologically rich nature of the Sinhala language, as well as the cultural biases of the English Wordnet as used in the PWN. The major needing resolution in the development process can be categorized as follows:

3.1 Morphological Forms

As mentioned above, Sinhala is a morphologically rich language which accounts for up to 110 noun word forms and up to 282 verb word forms. Therefore it is extremely important to incorporate a morphological parser to map such word forms to their corresponding lemmas. Table 1 shows some examples these morphological forms for moth nouns and verbs. A complete morphological parser for Sinhala is being developed at the Language Technology Research Laboratory (LTRL) of the UCSC and is expected to couple with the Sinhala wordnet to enhance the value of this resource.

Morph. Form	POS	Meaning	Lemma
බලමි *(baləmi)*	Verb	See (1st person, Sg)	බලනවා *(balənəvä)*
බැලීය *(bælïyə)*	Verb	See (3rd Person, Sg)	බලනවා *(balənəvä)*
බලද්දී *(baladdï)*	Verb	While Seeing	බලනවා *(balənəvä)*
බල්ලෝ *(ballö)*	Noun	Dog (Nominative, Pl)	බල්ලා *(ballä)*
බල්ලන් *(ballan)*	Noun	Dog (Accusative, Pl)	බල්ලා *(ballä)*
බල්ලාගේ *(ballägë)*	Noun	of Dog	බල්ලා *(ballä)*

Table 1. Morphologically different forms which share the same lemma

3.2 Compound Nouns and Verbs

Compounding is a very productive morphological process in Sinhala. Both Sinhala nouns and verbs formed by compounding nouns (nouns) and nouns with verbs (e.g. verbs *do* and *be*) are extremely productive. As a result of this compounding, the original sense of the constituents of the compound noun is altered, resulting in the derivation of a new sense. The methodology we used to extract the most important senses (as explained in 2.1) does not detect compound words, since we used the most frequent *single words* extracted from the corpus.

3.3 Language and Culture Specific Senses

Several culture specific senses were among the most frequent Sinhala words which had no corresponding sense IDs defined in PWN (e.g., "මිරිස් ගල" *miris galə* - "A flat stone and drum stone use to grind chilly, curry powder etc.", "පොල් ගෑනවා" *pol gänəvä* - "The act of scraping coconuts using a coconut scraper"). Two possible approaches were identified to find the appropriate place in the ontology for such senses. The first was to find the closest approximation in the existing ontology for an equivalent concept. The second was to extend the ontology appropriately to accommodate these concepts in order to represent them most accurately.

3.4 Word Selection Criteria

The words for the Sinhala wordnet were chosen from the UCSC Sinhala Corpus as described in Section 2.1. Many of these words have senses in standard Sinhala dictionaries that are not used in contemporary Sinhala. It was identified that taking these senses of words into account is not useful for the goals of the current project, and therefore they were ignored after carefully examining the period to which the usage of such senses belong.

4 Future Work

The process of building a Sinhala wordnet was mainly targeted as a resource for aiding language processing tasks. Hence aspects of providing an integrated GUI were not given priority and the resource stands on its own as a structured text document. It is expected to be integrated with a Sinhala morphological parser (which is currently being developed) in order to be of practical use. Therefore it is necessary to integrate this lexical resource with a comprehensive tool for manipulating data easily.

The current Sinhala Wordnet consists of 1,000 of the most common senses of contemporary Sinhala usage. Lexical relations of these words have been automatically linked to the English Wordnet due to adopting PWN sense IDs, even though some entities related to these 1,000 words are not present in English. Therefore it is essential to expand the Sinhala wordnet for these links and also to add senses according to importance, in order to build a comprehensive Sinhala lexical resource.

The AsianWordNet (AWN) Project of the TCLLab of NECTEC in Thailand is an initiative to interconnect wordnets of Asian languages to which the present Sinhala Wordnet is being linked. It is hoped that this effort will lead to a comprehensive multi-lingual language resourse for Asian languages.

5 Conclusion

Building a lexical resource such as wordnet is essential for language processing applications for the less resourced languages of the world. However the task requires significant resource allocations and expert knowledge to build for a particular language. As such, if a 'newly digitized' language can benefit from already developed linguistic infrastructure for another language, much effort can be saved. In the process of such adoption however, certain adaptations may need to be performed owing linguistic and cultural peculiarities of the language concerned.

This paper recommends the use of corpus statistics to identify the most important senses for a particular language to encode in a wordnet, in any given phased implementation effort. Such statistics provide a way to identify the most frequently used word senses specific to a culture which need to be dealt with first in order to get the highest return on investment of effort.

For languages which are morphologically rich, a morphological parser needs to be incorporated as a front end to such lexical resources. Many of the most frequent words of this kind of agglutinative language are irregular in form, requiring a morphological analyzer able to handle such forms.

Acknowledgements

This work was carried out under Phase 2 of the PAN localization project funded by IDRC of Canada. Authors acknowledge the contribution of several members of the Language Technology Research Laboratory of the University of Colombo of School of Computing in building the Sinhala Wordnet. In particular, the contribution of Vincent Halahakone in proofing Sinhala-English translations using his immense experiences as an English language teacher for government schools and universities is gratefully acknowledged. The authors also acknowledge the feedback given by two reviewers of this paper which helped in improving the quality of the work reported. Any remaining errors however, are those of the authors.

References

Chakrabarti, D. and Bhattacharyya, P. 2004, *Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT,* Global WordNet Conference (GWC-2004), Czech Republic.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya 2002, *An Experience in Building the Indo WordNet - a WordNet for Hindi*, First International Conference on Global WordNet, Mysore, India.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, P., Fellbaum, C. 2006. *Building a WordNet for Arabic.* LREC, Italy.

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

Putra, D.D, Arfan, A., Manurung R. 2010. *Building an IndonesianWordNet*, University Indonesia.

Roget, P.M , Roget, J. L. , Roget, S. R. 1962. *Thesaurus of English Words and Phrases*. Penguin Books.

Vossen, P. (ed) 2002. *Euro WordNet General Document*. Vrije Universiteit, Amsterdam

Wijayathunga, Harischandra 2003. *Maha Sinhala Sabdakoshaya*. M. D. Gunasena & Co. Ltd., Colombo