Building a WordNet for Sinhala

Indeewari Wijesiri
University of Moratuwa
Moratuwa, Sri Lanka
indeewari.wijesiri.09@cse.mrt.ac.lk

Malaka Gallage
University of Moratuwa
Moratuwa, Sri Lanka
malaka.gallage.09@cse.mrt.ac.lk

Buddhika Gunathilaka
University of Moratuwa
Moratuwa, Sri Lanka
buddhika.09@cse.mrt.ac.lk

Madhuranga Lakjeewa
University of Moratuwa
Moratuwa, Sri Lanka
lakjeewa.09@cse.mrt.ac.lk

Daya C. Wimalasuriya
University of Moratuwa
Moratuwa, Sri Lanka
chinthana@cse.mrt.ac.lk

Gihan Dias
University of Moratuwa
Moratuwa, Sri Lanka
gihan@uom.lk

Rohini Paranavithana
University of Colombo
Colombo, Sri Lanka

Nisansa de Silva
University of Moratuwa
Moratuwa, Sri Lanka
nisansadds@cse.mrt.ac.lk

Abstract

Sinhala is one of the official languages of Sri Lanka and is used by over 19 million people. It belongs to the Indo-Aryan branch of the Indo-European languages and its origins date back to at least 2000 years. It has developed into its current form over a long period of time with influences from a wide variety of languages including Tamil, Portuguese and English. As for any other language, a WordNet is extremely important for Sinhala to take it into the digital era. This paper is based on the project to develop a WordNet for Sinhala based on the English (Princeton) WordNet. It describes how we overcame the challenges in adding Sinhala specific characteristics which were deemed important by Sinhala language experts to the WordNet while keeping the structure of the original English WordNet. It also presents the details of the crowdsourcing system we developed as a part of the project - consisting of a NoSQL database in the backend and a web-based frontend. We conclude by discussing the possibility of adapting this architecture for other languages and the road ahead for the Sinhala WordNet and Sinhala NLP.

1 Introduction

Despite being used by over 19 million people and being one of the official languages of Sri Lanka, there has not been much progress in developing natural language processing (NLP) applications for the Sinhala language. This is partly due to the lack of commercial interest on developing Sinhala NLP applications on a global scale. For instance, as of now, neither Google Translate [1] nor Google News [2] is available for Sinhala while both are available in Hindi and Tamil – two other regional languages spoken by a much larger population and thus with a higher business value.

Within this backdrop, we believe that developing a fully functional WordNet for Sinhala would provide a much needed boost for the Sinahla NLP work. This is because it is well recognized that a WordNet is a very important tool in performing natural language processing tasks for any language. A WordNet will be helpful to Sinhala NLP application developers in tasks ranging from word sense disambiguation and information retrieval to translation. Moreover a Sinhala WordNet will be a valuable resource to linguists

[1] http://translate.google.com/
[2] https://support.google.com/news/answer/40237

studying the Sinhala language. We paid special attention to the interests and concerns of the latter group as described later in the paper.

The project team, mainly consisting of personnel from the Knowledge and Language Engineering Lab of University of Moratuwa, started the task of developing a WordNet for Sinhala with several brainstorming sessions which involved Sinhala language experts, computer science specialists and people who had previously made some contributions in digitizing the Sinhala language (for example in developing Sinhala Unicode characters). Although we were biased towards using the expansion approach, which develops a WordNet based on an existing WordNet for another language, we discussed the possibility of adopting the merge approach, which develops a WordNet using the first principles by leveraging existing dictionaries and other resources (Bhattacharyya, 2010). We settled on the expansion approach because it was evident that we do not have the resources to successfully pursue the merge approach.

We came up with basic design for the WordNet through the above mentioned brainstorming sessions and then proceeded to develop the technical infrastructure needed. This consists of developing Sinhala WordNet APIs and a web interface as well as a crowdsourcing system to add synsets and relationships. The latter is needed because coming up with Sinhala synsets and relationships based on the synsets of another language requires a lot of manual work. Initially we were planning to use the Hindi WordNet as the source WordNet but switched to the English WordNet a couple of months into the project. The reasons for this change are discussed in Section 2.2. Apart from this the development effort proceeded fairly smoothly and we have completed the implementation of the WordNet API and the crowdsourcing system. Currently we are in the process of adding synsets using this system.

The rest of the paper is organized as follows. In Section 2, we present the details of the discussions we had with Sinhala language experts and the effects these discussions had in the structure of the Sinhala WordNet. In Section 3 we discuss the technical details of the project. Here, we describe the use of a NoSQL database to facilitate modification to a WordNet, which has not been done before to the best of our knowledge. In Section 4, we describe how the crowdsourcing system works including how it gives suggestions to the contributors simplifying their task. We reflect on some important aspects of the project including the possibility of adopting the entire system to other languages in Section 5. We present the details of some related work in Section 6 and provide concluding remarks in Section 7.

2 Developing the Linguistic Infrastructure

Development of linguistic infrastructure was carried out as the first phase of the project. Several discussions with Sinhala language experts were conducted to better understand the key features of the Sinhala language.

2.1 Discussions with Sinhala Linguists

From the beginning of the project the development team was collaborating with some prominent experts on Sinhala language. The basic idea of this collaboration was to acquire the necessary knowledge of the Sinhala language to get to know the linguistic requirements of a Sinhala WordNet and to form an expert evaluator panel to help with the crowdsourcing effort in developing the WordNet.

One important topic discussed with the experts was that Sinhala has a significant difference in written and spoken usage. These differences include differences in word usage and differences in grammar. We were particularly interested in differences in word usage in spoken and written forms as grammar rules fall outside the scope of a WordNet. It was observed that words with subtle but important differences are used in the written and spoken forms of Sinhala. For instance, for the sense "man", මිනිසා (*minisa*) is the most frequent word used in written Sinhalese while මිනිහා (*miniha*) is the most frequent word used in spoken Sinhalese. While the difference is subtle (a single phoneme in this case) its implications are significant for a natural speaker of Sinhala. In this case, using මිනිසා in normal conversations appears extremely odd. Moreover such differences are very common and combining words used in spoken and written Sinhala results in very odd phrases.

The problem faced by us was whether to include this difference in the Sinhala WordNet. Doing so would go against the main objective of a WordNet which is organizing words by their meanings; clearly there is no difference in the meanings of මිනිසා and මිනිහා as it is simply a matter of language usage. Despite this concern, we decided to include this difference as a *flag* for each word due to the following reasons.

1. Not including these in the WordNet would result in the loss of a valuable opportunity to encode these differences in a machine readable manner; the contributors of the crowdsourcing system can do this with little extra effort but doing it as a separate project would require a lot more effort. The importance of this factor is magnified by the lack of commercial interest in Sinhala NLP.
2. Since one of the primary reasons for developing a Sinhala WordNet was to serve the needs of Sinhala linguists we wanted to accommodate their requirements. We suspected that eliminating this type of information would make the WordNet less useful to them. Janssen (2002) has made a similar argument with regards to eliminating gender information from WordNets. Hence, adding this information to the WordNet was seen as a pragmatic move.
3. Different words being used in spoken and written Sinhala is an extremely common phenomenon that cannot simply be ignored or left for later consideration.

By the same reasoning, we decided to add few more features of the Sinhala language to the WordNet. One of them is the gender difference. The genders in Sinhala are masculine and feminine but none are specified for some words (typically for things that are not alive). The gender of a noun is important as it decides which morphological form of a verb is used with it. Thus the Sinhala WordNet will contain the gender of each noun, if exists.

The Sinhala words can be divided into three main categories called native words, words directly borrowed from another language which are being used without any change (තත්සම - *tatsama*) and the words borrowed from another language and have been modified (තත්භව - *tatbawa*). The words have been mainly borrowed from Sanskrit, Pali, Hindi, Portuguese, English, Tamil and Dutch. In constructing phrases in Sinhala, the origin of the word should be considered similar to how the spoken/written differentiation is used. As an example 'mathru'(මාතෘ) and 'maw'(මව්) are two forms to express the meaning "mother's" in Sinhala but 'mathru' is a tatsama while 'maw' is a tatbawa. 'snehaya'(ස්නේහය) and 'senehasa'(සෙනෙහස) means 'affection' which again are tatsama and tatbawa. To express "mother's affection", people use either 'mathru snehaya'(මාතෘ ස්නේහය) or 'maw senehasa'(මව් සෙනෙහස) while the other two combinations appear odd. This is despite the fact that all four words are acceptable in written Sinhala. Thus details of the origin of a word are also included in the Sinhala WordNet. Both the source language and the derivation type (tatsama/tatbawa) are kept on this regard.

Each noun in Sinhala can be in 9 morphological forms called 'vibhakthi'(විභක්ති). Furthermore there are fairly complicated rules in forming compound words called 'sandi'(සන්ධි) and 'samasa'(සමාස). The formation of these forms and rules as well as the inflectional forms of a verb are based on the *root* of the word, which may not be the *most commonly used form* of the word. Therefore, it was decided to keep the *word root* as well as the *most common morphological form* in storing a word in the WordNet.

In summary, we decided to include the following features for each word.

- Written/ Spoken usage
- Gender
- Origin of the word
- Word root
- The most common morphological form

It is interesting to relate these features, which are deemed important in representing Sinhala words in a machine-processable format, to a standard lexical-encoding framework. Our discussion on this regards is based on the lemon (Lexicon Model for Ontologies) framework (McCrae et al., 2012). Our view is that the written/spoken usage and the origin of the word are properties under the *linguistic description module* of lemon outside its *core*. These will be used by the *phrase-structure module* in identifying well-formed phrases. The word root is related to the *morphology module* and is used in inflection while the most common morphological form is the main lexical entry in the *core* for the word in concern. The gender information is useful for inflection in the *morphology module* and in recognizing words that do not have certain morphological forms. (e.g., රැජිණ - rajina - the queen does not have a masculine form).

2.2 Selecting the Source WordNet

As mentioned earlier we decided to develop the Sinhala WordNet following the expansion approach due to practical considerations. Then the question was which WordNet to use as the source WordNet. We first decided to use the Hindi WordNet (Jha et al., 2001) for this purpose due to the following reasons.

1. The Sinhala language belongs to the Indo-Aryan branch of the Indo-European languages and is heavily influenced by the classical Indian languages of Sanskrit and Pali. Since Hindi is close to Sanskrit and the Hindi WordNet is fairly sophisticated - it serves as the hub of the Indo WordNet initiative (Bhattacharyya, 2010) - we assumed that the Hindi WordNet would provide a good basis for developing the Sinhala WordNet. We even considered using the Sankrit WordNet as the source WordNet but realized that it is still in an early stage.
2. The success of the Indo WordNet initiative in creating WordNets for many languages in India (Bhattacharyya, 2010) was one of the main motivations for us in embarking on this project. It was assumed that using the Hindi WordNet as the source WordNet would help us leverage the success of the Indo WordNet.

However, as we proceeded with the development work, it was apparent that using the Hindi WordNet as the source WordNet was not a viable option. The following are the main reasons for this.

1. Despite the perceived similarity in the origins of the languages, Hindi and Sinhala are very different languages in many aspects related to WordNet construction: One difficulty associated with this is that Hindi is written in Devanagari script, which is not familiar to most Sinhala speakers. (Sinhala has its own alphabet). Moreover, for many Hindi words it was difficult to identify Sinhala words with the same meaning, even after knowing how the word is pronounced. It was thought that translating Hindi words to Sinhala would be easier once the pronunciation is known because words of the languages are often pronounced similarly – e.g., Sinhala බෑයා (*baaya*) vs. Hindi भाई (*bhai*) meaning brother. It was seen that such similarities are not very common. As a result, we found ourselves frequently translating words from Hindi to English to understand the relevant Sinhala words.
2. It was seen that adopting the technical infrastructure of the Indo WordNet project to develop the Sinhala WordNet was difficult. Part of this is due the communication difficulties – all other WordNets of the Indo WordNet have been developed within India itself. In addition, our requirement to add *flags to words* in addition to *flags for synsets* as described in Section 2.1 created additional complexities and we found that accommodating these changes in the Indo WordNet text database stricture was very difficult. The Princeton English WordNet (Fellbaum, 1998), with its extensive documentation and the support network was seen as a much better alternative in this context.
3. A significant percentage of native Sinhala speakers have a working knowledge in English and it was seen that this will be very useful for a crowdsourcing system. In contrast, familiarity with the Hindi language is not widespread and this combined with the fact that most Hindi words are apparently unfamiliar to Sinhala speakers as described in (1), means that it is very difficult to use the Hindi WordNet in a crowdsourcing system.

Based mainly on the above factors, we switched the source WordNet from Hindi to English early in the development stage. The fact that the WordNets for Arabic (Rodrıguez et al., 2008) and Japanese (Isahara et al., 2008), which have very little in common with English, have also been developed with the English WordNet as the source, also weighed in on our decision.

We were mindful of the consequences of using the English WordNet as the source WordNet in developing the Sinhala WordNet. It has been stated that the source WordNet can have a distracting influence on the new WordNet being created especially when the two languages exist in different regions and cultural settings (Bhattacharyya, 2010). It is clear that this concern is applicable here. As such we decided to aggressively remove existing synsets in the English WordNet and add new synsets as necessary when developing the Sinhala WordNet.

3 Developing the Technical Infrastructure

After developing the linguistic infrastructure, we focused on developing the technical infrastructure according to the requirements identified. The main challenges we faced here were resolving the complications arising when extending the Princeton WordNet API, dealing with different data structures, and selecting tools and technologies. In this section, we describe the salient features of the architecture of the system and how we approached the above mentioned challenges.

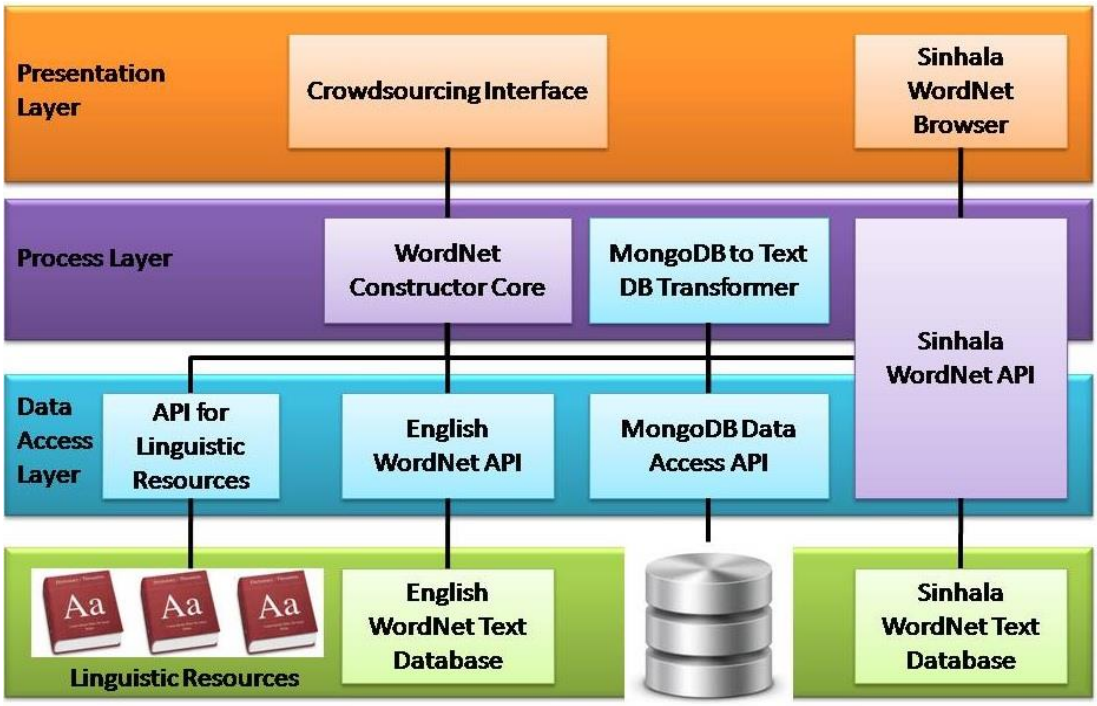

Figure 1: System Architecture

3.1 The WordNet API

The Sinhala WordNet API is implemented on the Java platform extending the English WordNet API (JWNL)[3]. The basic idea of developing this API is to provide general WordNet functionalities as well as the specific functionalities of the Sinhala WordNet discussed above. We defined new classes for synset, word, noun, verb, adjective and adverb extending the JWNL classes. The JWNL documentation and mailing lists were extremely helpful to us in this exercise. Incorporating Sinhala characters in the API was based on the Sinhala Unicode characters.

3.2 System Architecture

Figure 1 shows the architecture of the entire system, consisting of the API and the crowdsourcing system. For the non-technical users, the main outputs of the system are the online and offline Sinhala WordNet browsers and the web-based interface for the crowdsourcing system. Developers will have access to these components as well as the source code of the Sinhala WordNet API, WordNet Constructor Core - which governs how the crowdsourcing system operates -, the MongoDBToTextDB Transformer and the schemata of the underlying databases.

[3] http://jwordnet.sourceforge.net/handbook.html

The components in the presentation layer get the data they need from three sources.

1. The English WordNet: The data contained in the English WordNet text database in terms of synsets and relationships are used.
2. The NoSQL Database: The modifications made by contributors of the crowdsourcing system to the data of the English WordNet are stored in this database.
3. Linguistic Resources: Several linguistic resources such as available machine readable dictionaries for Sinhala are used in providing suggestions for the collaborators.

Components in the Data Access Layer are used by the two components in the Process Layer to access the necessary data.

The MongoDBToTextDB transformer gets the data from the NoSQL database as well as the text database of the English WordNet because the NoSQL database *only* contains the modifications made by collaborators. It combines the data from the two sources into the text database of the Sinhala WordNet API. This step is carried out when releasing a new version of the Sinhala WordNet.

3.3 Use of a NoSQL Database

According to the system architecture described above, we need a database to store the modifications performed by the contributors of the crowdsourcing system. The modifications include adding Sinhala words to a synset, adding features to words and synsets, adding relation-

ships between words/synsets and adding and removing synsets.

Until recently, the standard solution for this type of a data storage need has been to use a relational database system. However, the use of NoSQL databases has increased in the recent past partly due to the flexibility it offers to the schema designer. Instead of being restricted to a relational schema, which often requires multiple tuples spread across several relations for the same logical data unit, NoSQL databases allows the designers to store data according to the semantics behind them. We realized that these advantages will be important in our system since a synset consists of an unlimited number of words, each with several distinct features.

Another advantage of using NoSQL databases is that they provide better scalability than relational database systems especially in setting up multiple servers connected to a web-based front-end. This too will be helpful in using a crowsourcing approach for WordNet creation as the system will provide better performance for the contributors.

Noun
_id
_class
userName
EWNID
Words
_id
Lemma
wordID
wordPointerList
pointerType
synsetType
synsetId
wordId
sensePointers
pointerType
synsetType
synsetId
gloss

Table 1: Schema for Nouns

However, it was noted that NoSQL solutions *do not* guarantee consistency of the database although they provide *eventual consistency*. Therefore, it is possible, in rare conditions, for two contributors to make contradictory updates in the database. In the context of our system, these inconsistencies can be resolved later, generally in evaluation. Moreover any inconsistencies do not affect the releases of the Sinhala WordNet as they use the text database, assuming that any contradictions are resolved before a release.

We concluded that the advantages of NoSQL databases outweigh their disadvantages and decided to use one. We selected the MongoDB NoSQL (Plugge et al, 2010) system. Table 1 shows the schema we used for nouns. To the best of our knowledge, this is the first time a NoSQL database has been used in developing a WordNet.

Currently, the source repository is maintained as a private GitHub project. We will make it public in the near future.

4 The Crowdsourcing System

4.1 Overview

As mentioned earlier, a crowdsourcing system to facilitate the development of the Sinhala WordNet was designed and implemented as a part of the project. As illustrated in Figure 1, the WordNet Constructor Core component contains the major functionalities of this system. It obtains different types of data through the components of the Data Access Layer and provides an interface to be used by the web-based interface of the crowdsourcing system. The following are the different types of data used by this component through the Data Access Layer.

1. Information contained in the English WordNet through the EWN API (JWNL).
2. Information obtained from several linguistic resources for the Sinhala language including machine readable dictionaries and thesauri. These are used to specify suggestions to contributors to simplify their task as described in Section 4.2.
3. Information contained in the mongoDB database, which contains the modifications made by the contributors as mentioned earlier.

The web-based user interface allows contributors to browse through the English WordNet hierarchy and perform modifications as necessary. If no work has been done on a particular synset of the English WordNet, they will be shown the data contained in the English WordNet and are expected to replace them with Sinhala words. These changes include adding words to synsets, specifying flags for the words (e.g., whether the word is used in written/spoken Sinhala) and adding relationships. All the modifications are saved in the MongoDB database.

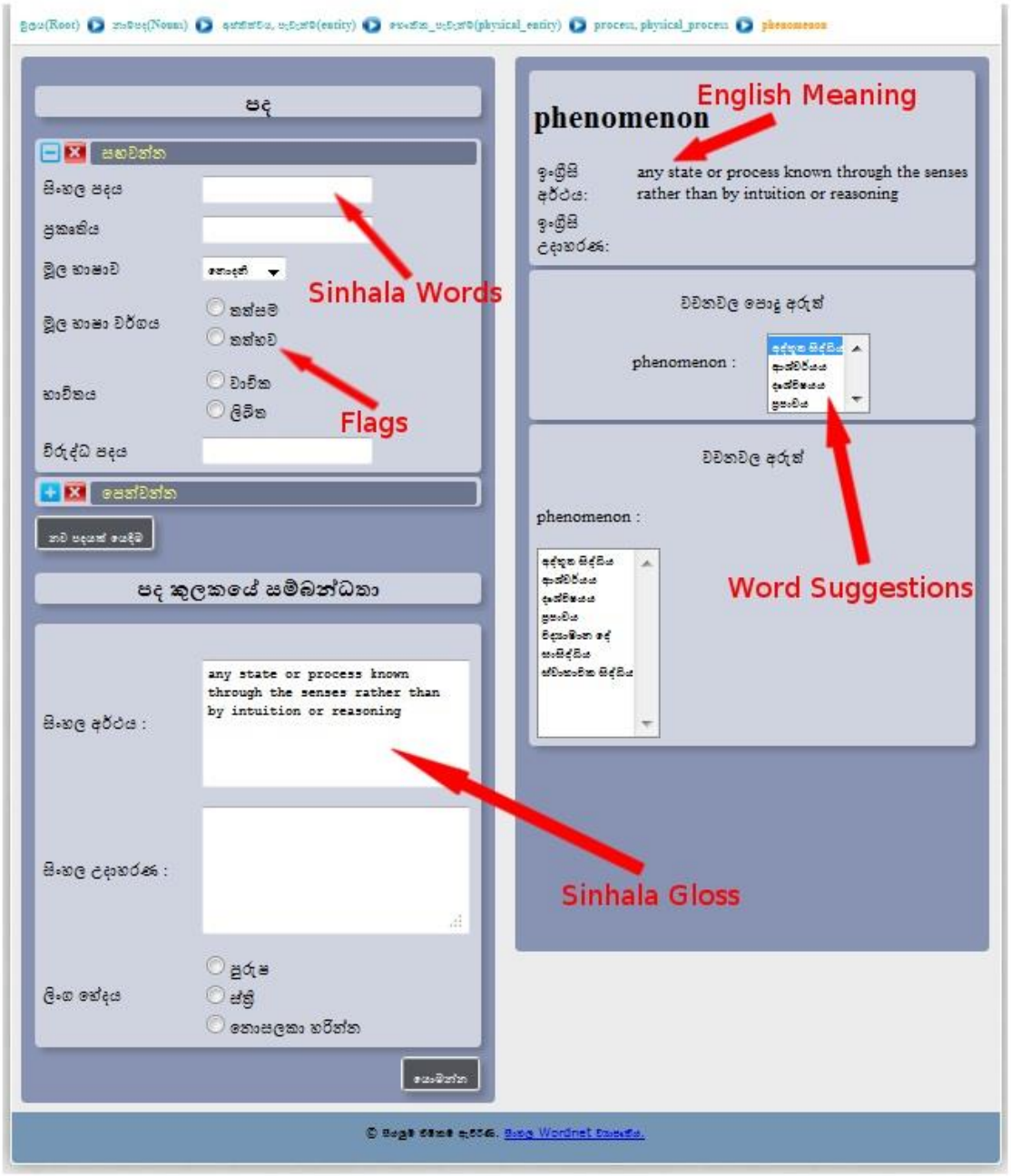

Figure 2: The UI of the Crowdsourcing System

Figure 2 shows the web interface when adding Sinhala words/relationships for the English synset for one sense of the word "phenomenon". Since Sinhala words have not been added to this synset, it shows the available information in the English WordNet. In addition, it shows suggested Sinhala words obtained from linguistic resources as described in Section 4.2.

The web-based user interface is operational and can be accessed from `http://www.wordnet.lk`. The modifications made by the contributors have to be approved by an evaluator before being included in a release.

How to effectively use a crowdsourcing technique to get a particular task done with acceptable quality is an open research question. Dow et al. (2012) have found that assessment of work produced, whether it is external assessment or self-assessment, if very helpful on this regard. As such, we expect the feedback provided by evaluators to help our effort.

4.2 Providing Suggestions

The purpose of providing suggestions for contributors is simplifying their task so that they do not have to rely entirely on their knowledge and available printed material. Currently, we provide suggestions for English words based on machine readable English to Sinhala and Sinhala to Sinhala dictionaries and thesauri. Out of the available resources, we found the Madura English-Sinhala dictionary (Kulatunga, undefined) particularly helpful. We are currently in the process of improving this component by incorporating the the-

sauri developed by the Department of Official Languages of Sri Lanka and a text corpus compiled by ourselves.

5 Discussion

5.1 The Morphology of the Language

Sinhala is an inflectional language where many verbs and nouns have a fairly large number of morphological forms. Verbs and nouns frequently have more than 10 morphological forms when considering both spoken and written forms. This has implications for the WordNet as a person or a software system searching for a word may use a different morphological form from what is contained in the WordNet. We decided against storing all morphological forms of a word in the WordNet since that increases the number of words for a synset to an unmanageable level. As such a good morphological analyzer, which is external to the WordNet is necessary to obtain the full benefits of the WordNet. There have been previous attempts to develop a morphological analyzer for Sinhala which have produced satisfactory results (Hettiage, 2006; Fernando and Weerasinghe 2013).

5.2 Extending to Other Languages

While we did not develop our system with the objective of developing WordNets for languages other than Sinhala, we recognize that it has the potential to be used in this manner. The architecture of the system has to be changed in some places, for example in using linguistic resources of other languages for providing suggestions for contributors. But the overall design of displaying the information of the English WordNet, allowing the contributors to modify them with words from the target language and storing the modifications in the NoSQL database can be easily applied in developing a WordNet for another language based on the English WordNet following the expansion approach. It is possible to reuse the schema of the MongoDB database and the source code of the crowdsourcing interface, the WordNet Constructor Core and the MongoDBToTextDB Transformer in such an exercise. We plan to separate out these parts from our codebase as a future work.

5.3 Current Status

The crowdsourcing system is currently operational and the number of synsets in the Sinhala WordNet is approaching 2000. This number is significant since this has been used as a marker by the Indo WordNet project in developing WordNets for languages in India (Bhattacharyya, 2010). Our goal is to release the first complete version early next year.

The Knowledge and Language Engineering Lab of the Department of Computer Science and Engineering at University of Moratuwa is coordinating this effort.

6 Related Work

The Hindi WordNet and the Indo WordNet initiative provided a lot of inspiration to us in attempting to develop a WordNet for Sinhala following the expansion approach. We followed their work in several aspects of the project such as the use of crowdsourcing to generate synsets.

There has been a previous work on developing a WordNet for Sinhla by Welgama et al. (2011), which is basically an exploration on developing a WordNet for Sinhala by extracting some common words from a corpus and getting the help of Sinhala language experts to come up with synsets based on them. It can be seen that this work is related to the merge approach. Our work differs from this effort in our use of the expansion approach and the objective of developing a complete WordNet.

7 Conclusion

Developing a fully functional Sinhala WordNet can be considered a landmark in NLP for Sinhala and we believe that we are well set to achieve this in the near future. This will provide a tremendous boost for developing Sinhala NLP applications such as information retrieval systems, text classifiers and summarizers and translators. The availability of a platform in terms of a WordNet may even attract some commercial interest for Sinhala NLP.

It should also be recognized that our work has the potential to be generalized into a system that can be used to bootstrap WordNet creation for a language. If this goal can be achieved, it will be extremely helpful in developing WordNets for minority languages such as Sinhala.

Acknowledgements

We thank Prof. J.B. Disanayaka, Dr. Sandagomi Coperahewa and Mr. Achinthya Bandara of the Department of Sinhala of University of Colombo and Mr. Anushke Guneratne of the LK Domain Registry for their help in this project.

References

Pushpak Bhattacharyya. 2010. IndoWordNet, *Proceedings of the Lexical Resources Engineering Conference*.

Steven P. Dow, Anand Kulkarni, Scott R. Klemmer and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work, *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*: 1013-1022.

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

Niroshinie Fernando and Ruwan Weerasinghe. 2013. A Morphological Parser for Sinhala Verbs. *Proceedings of the International Conference on Advances in ICT for Emerging Regions.*

Buddhita Hettige. 2006. A Morphological Analyzer to Enable English to Sinhala Machine Translation, *Proceeding of the 2nd International Conference on Information and Automation*: 21-26.

Hitoshi Isahara, Fransis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Maarten Janssen. 2002, Differentiae Specificae in EuroWordNet and SIMuLLDA. *Proceedings of the Ontologies and Lexical Knowledge Bases Workshop*.

Madura Kulatunga. (undefined). Madura English-Sinhala Dictionary. Retrieved September 6, 2013, from `http://maduraonline.com/`.

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura A Hollink, Elena Montiel-Ponsoda, Dennis Spohr and Tobias Wunner. 2012. *Interchanging lexical resources on the Semantic Web, Language Resources and Evaluation, 46(4)*: 701-719.

S. Jha, Dipak Narayan, Prabhakar Pande and Pushpak Bhattacharyya. 2001. A WordNet for Hindi, *Proceedings of the International Workshop on Lexical Resources in Natural Language Processing*.

Eelco Plugge, Tim Hawkins and Peter Membrey. 2010. *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*. Apress.

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Antonia Martí, William Black, Sabri Elkateb , James Kirk, Piek Vossen and Christiane Fellbaum. 2008. Arabic WordNet: Current state and future extensions. *Proceedings of the Fourth Global WordNet Conference*.

Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe and Tissa Jayawardana. 2011. Towards a Sinhala WordNet, *Proceedings of the Conference on Human Language Technology for Development*.