Sinhala Short Sentence Similarity Measures using Corpus-Based Similarity for Short Answer Grading

JCS Kadupitiya, Surangika Ranathunga, Gihan Dias
Department of Computer Science and Engineering
University of Moratuwa, Katubedda 10400
Sri Lanka
{jcskadupitiya.16,surangika,gihan}@cse.mrt.ac.lk

Abstract

Currently, corpus based-similarity, string-based similarity, and knowledge-based similarity techniques are used to compare short phrases. However, no work has been conducted on the similarity of phrases in Sinhala language. In this paper, we present a hybrid methodology to compute the similarity between two Sinhala sentences using a Semantic Similarity Measurement technique (corpus-based similarity measurement plus knowledge-based similarity measurement) that makes use of word order information. Since Sinhala WordNet is still under construction, we used lexical resources in performing this semantic similarity calculation. Evaluation using 4000 sentence pairs yielded an average MSE of **0.145** and a Pearson correlation factor of **0.832**.

1 Introduction

There has been no research conducted for measuring similarity between short sentences written in Sinhala, an official language of Sri Lanka, which is currently used by a population of over 16 million.

Several unsupervised techniques are used for short sentence similarity calculations. These unsupervised approaches can be categorized in to four basic classes: corpus-based, knowledge-based, string-based, and other similarity measures (e.g. those that consider word order and word length). Corpus-based similarity determines the similarity between two sentences/texts according to information gained from a corpus. Knowledge-based similarity measures are based on identifying the degree of similarity between words using information derived from semantic networks (e.g. WordNet) or lexical resources. Corpus-based and knowledge-based measures are also referred to as semantic similarity measures (Li, 2006). String-based similarity measures operate on string sequences and character composition. This technique can be further divided in to character-based similarity measures and term-based similarity measures. Even though each of these techniques could be directly used to calculate the similarity of two given sentences, much previous research work combined two or more approaches to form hybrid similarity measuring techniques to gain a higher accuracy (Li, 2006; Zhao, 2014). The most popular hybrid techniques include corpus based similarity calculations, and knowledge based similarity calculations that use WordNet for Word Sense Disambiguation (WSD). For English, the most promising results were given by the latter. The former technique does not require special Natural Language processing (NLP) tools other than a corpus. In contrast, the latter requires many NLP resources such as part of speech (POS) taggers, lexical databases, word lists, and corpora in addition to WordNet. However, as an under-resourced language, development of many of these basic resources for Sinhala is still at inception stage (Welgama, 2011; Weerasinghe, 2013).

This research focuses on finding the best possible NLP technique(s) for similarity calculation between short Sinhala phrases by utilising existing unsupervised techniques for English. Constrained by the available resources, we experimented with two hybrid techniques: semantic similarity measures that make use of word order information as presented by Li et. al's (2006), and semantic similarity measures that make use of word length information as presented by Zhao (2014). Both these hybrid similarity measures make use of corpus based and knowledge based approaches plus a basic lexical database, and

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 44–53, Osaka, Japan, December 11-17 2016.

domain-specific word glossaries. Best results were given for the first approach that made use of word order information. The rest of the paper is organized as follows. Section 2 discusses previous work on short sentence similarity in general. Section 3 provides the methodology whereas section 4 describes the results and discussion. Conclusion and limitations of the current implementation, and suggestions for future work are given in sections 5 and 6, respectively.

2 Related work

Techniques for short sentence similarity measurement can be broadly categorised into two groups as unsupervised and supervised approaches. In this section, we only discuss unsupervised techniques, as this is what is employed in our research. However, we mention in passing that most of the methodologies used in supervised approaches require WordNet, morphological analyser, and/or a POS tagger to generate the features (Mohler, 2011; Alves, Bestgen, Biçici and, Zhao 2014), whereas most of the unsupervised approaches do not require these resources. Moreover, as reported by some researchers, unsupervised techniques have performed well than supervised approaches in some situations (Marelli, 2014).

As mentioned earlier, previous research focused on combining two or more unsupervised approaches to form a hybrid similarity measuring technique to gain a higher accuracy.

Gomaa (2012) employed thirteen well-known algorithms (Damerau-Levenshtein, Jaro, Jaro–Winkler, N-gram, Cosine Similarity, etc.) to calculate the similarity score between two short English sentences. Six of these algorithms are character-based and the other seven are term-based measures. For the corpus based similarity measures they have used Latent Semantic Analysis (LSA) and Explicit Semantic Analysis (ESA). Gomaa (2012) claims that the best results are given when N-gram was combined with LSA.

A research focused on similarity calculation for Hindi language employs knowledge-based similarity approaches using WordNet and String-Based approaches (Tayal, 2014). They claim that semantic similarity calculation can be applied for any Indic language such as Hindi, Marathi. Sinhala also belongs to this branch of the language tree.

Mohler et. al (2009) has done a comprehensive evaluation of different knowledge-based and corpus-based measures for the task of short answer grading using both corpus-based algorithms and knowledge-based algorithms. Their techniques make use of WordNet hierarchy and Wikipedia corpus. They conducted comparative evaluations using eight knowledge-based measures of semantic similarity (shortest path, Leacock and Chodorow(1998), Lesk(1986), Wu & Palmer (1994), Resnik (1995), Lin (1998), Jiang & Conrath (1997), Hirst and St-Onge, (1998)), and two corpus-based measures (LSA and ESA) . Out of all these techniques, the best results were given for the LSA approach.

A research done by Li et. al's (2006) focused on sentence similarity measurement based on a hybrid approach by combining semantic similarity measures (knowledge and corpus based similarity measures) and, word order based similarity measures. It presents an algorithm that takes account of semantic information and word order information implied in the sentences. The semantic similarity of two sentences is calculated using information from the WordNet and from the corpus statistics using Brown corpus. In this approach, a sentence is considered as a sequence of words, each of which carries useful information about the meaning. The words and their combined structure make a sentence to convey a particular meaning. When comparing all the possible unsupervised techniques for English short sentence similarity, Li's (2006) method has given the most accurate results.

Recent research work done by Zhao (2014) has focused on a combined unsupervised approach using knowledge based similarity measures (8 similarity measures based on WordNet : Wu & Palmer (Wu and Palmer, 1994), Resnik (Resnik, 1995), etc) and word length based similarity measurements (8 similarity measures, which are further described in section 3.3). They have combined knowledge based feature vector and length measure vector for their final similarity calculation. This has outperformed author's supervised approach for the similarity calculation task.

3 Methodology

As described in the literature review, most of the unsupervised techniques do not require much NLP resources, and the techniques are language independent to a great extent. Moreover, unsupervised techniques have given comparable, or even better results than supervised approaches in some cases. Due to these facts, we decided to follow an unsupervised approach in this research.

We identified that Li's (2006) methodology has given the best results among other research for semantic similarity based techniques we referred to ((Gomaa, 2012) and (Mohler, 2011)). This approach focuses on combining semantic similarity measures (knowledge-based and, corpus based similarity measures) and word order based similarity measures to form a hybrid approach. In the absence of Sinhala WordNet, we modified Li's (2006) knowledge-based similarity measures to use the Sinhala lexical resources we created considering similar word sets. We also modified his corpus based similarity calculation methodology to consider statistical information taken from Sinhala word glossaries. Other than this, Li's (2006) methodology is language-independent.

Following Zhao (2014), we also tried combining semantic similarity calculation with word length based similarity measures, however, this did not outperform our previous approach.

3.1 Data Preparation

In the Semeval-2014 task 1[1], a dataset called SICK was built using the 8K ImageFlickr[2] data set (Marelli, 2014). The SICK data set consists of about 10,000 English sentence pairs, each sentence pair was annotated for relatedness and entailment by means of crowdsourcing techniques. Similar to the approach followed for data set preparation in this task, we selected 500 images from this dataset and asked five participants to describe each image using one short Sinhala sentence. Thereby we collected 2500 short Sinhala sentences. We randomly formed 5000 sentence pairs from these 2500 sentences. Finally, we employed another three persons to manually annotate these pairs with a score from 0 to 5 (with 0 being completely dissimilar and 5 being exactly similar). Table 1 shows example sentence pairs with different degrees of semantic relatedness; gold relatedness scores are expressed on a 6-point rating scale. For the final evaluation, these scores (between 0-5) were normalized to form a similarity score that lies between 0 and 1.

Relatedness score	Example Sentence Pair
3.34	A : මිනිසෙකු වාහනයක් අලුත් වැඩියා කරයි (A man repairs a vehicle)[3] B : මිනිසෙක් ඔසවා ඇති මෝටර් රථයක් සෝදමින් සිටියි (A man is washing a motor car, which is lifted)
2.34	A : මිනිසෙක් යතුරු පැදියක් ධාවනය කරයි (A man rides a motorcycle) B : යතුරු පැදි ධාවකයෙක් දකුණට වංගුවක හැරෙයි (A motorcycle rider takes a right turn at a bend)
3.67	A : තරඟයක ක්‍රීඩකයෝ තිදෙනෙක් ගුවනේ ඇති පන්දුව ග්‍රහණය කරගැනීමට පොර කති (In a game, three players are competing to grab the ball that is in the air) B : පාපන්දු ක්‍රීඩකයෙක් තවත් ක්‍රීඩකයෙකුගෙන් පන්දුව ලබා ගැනීමට උත්සහ කරයි (A Football player tries to get the ball from another player)
0.00	A : ක්‍රීඩකයෙක් අශ්වයාගේ පිටින් වැටෙයි (A player falls from a horseback) B : බේස්බෝල් ක්‍රීඩකයෙක් කලු පිත්තක් අතින් අල්ලාගෙන සිටියි (A baseball player is holding a black bat by the hand)

Table 1: Example sentence pairs with their gold relatedness scores (on 6-point rating scale).

3.2 Sinhala Lexical Database and Domain Specific Glossaries

Almost all the knowledge-based techniques reviewed in section 2 employ WordNet for calculating semantic similarity between short sentences (Li, 2006; Mohler, 2009; Tayal, 2014; Zhao, 2014). However, WordNet for Sinhala[4] is still under construction (Welgama, 2011; Wijesiri, 2014). Thus we opted to use a Sinhala lexical database, as approaches that employed lexical databases have also given performance results similar to those employed WordNet (Corley, 2005). Accordingly, we created a Sinhala lexical database consisting of 195781 words and 30564 synsets using online dictionaries (English-Sinhala). This lexical resource is created in such a way that all the words similar in meaning share a unique identification number. Using our lexical resource, we were able to check whether two Sinhala words are

[1] http://alt.qcri.org/semeval2014/task1/

[2] http://nlp.cs.illinois.edu/HockenmaierGroup/data.html

[3] Each Sinhala sentence was manually translated to English by the author, so that a wider audience can understand.

[4] http://ucsc.cmb.ac.lk/ltrl/?page=panl10n_p2&lang=en

similar or dissimilar, but we are unable to get partial relatedness values as given by WordNet synsets. We also used domain specific word glossaries from the Department of Official Languages[5], Sri Lanka. These glossaries[6] are for 22 domains such as education, statistics, physics, mathematics, sports, and linguistics.

3.3 Semantic Similarity Calculation using Word Order Information

Fig. 1 shows the procedure for calculating the semantic similarity between two candidate sentences using the technique presented by Li et. al's (2006). In this approach, a vector is dynamically formed in the form of a Bag of Word vector (BoW vector) considering the occurrence of unique words in the two sentences. For both sentences (S_1 and S_2), raw vectors (v_1 and v_2) are derived with the help of the lexical resources. Each entry in the raw vector corresponds to a word in the BoW, so the dimension of the vectors equals the number of unique words in the two sentences. When creating the raw vectors, we consider two cases: if word appears in the sentence, corresponding element of the vector is set to 1, if word does not appear in the sentence, lexical resources are used to check whether a similar word is there. If it is there, corresponding element of the vector is set to 1 and if it is not there, vector element is set to 0. Then v_1 and v_2 are further processed to form two semantic vectors (V_1 and V_2). Here, since every word in a sentence differently contributes to the meaning of the whole sentence according to the domain in which we compare the similarity, a weight is introduced to the words. This weight is introduced as the TF-IDF (term frequency-inverse document frequency) value for the particular word considering relevant domain specific glossary vs. other available glossaries. Sports domain glossary is selected as specific glossary as our dataset was created using mostly sports images. Semantic similarity between two sentences ($S_{1,2}$) is defined as the cosine coefficient between the two vectors V_1 and V_2.

As in other comparable Indic languages (e.g. Hindi), stop words in Sinhala sentences also carry very important information about the semantic similarity (Tayal, 2014). Because of that, we chose not to remove stop words.

Now consider the below sentences, S_1 and S_2.

S_1 : මිනිසෙකු බල්ලෙකු මතට සතුටින් පනී (A man happily jumps onto a dog)
S_2 : බල්ලෙකු මිනිසෙකු මතට සතුටින් පනී (A dog happily jumps onto a man)

If the two sentences (S_1 and S_2) contain the same set of words, any method based on the BoW model will give a decision that S_1 and S_2 are exactly the same.

However, it is clear to a human eye that S_1 and S_2 are not same. The dissimilarity between S_1 and S_2 is due to the word order. Therefore, the similarity calculation method for sentence comparison should consider the impact of word order as well.

The right hand side of Fig. 1 shows the procedure for calculating the word order similarity between two candidate sentences. For the sentence pair S_1 and S_2, the joint word set ($S = S_1 U\, S_2$) can be formed as:

S : { මිනිසෙකු, බල්ලෙකු, මතට, සතුටින්, පනී } (a man, a dog, onto, happily, jumps)

If we assign a unique index number for each word in S_1 and S_2, we can form two word order vectors (r_1 and r_2). The index number is simply the order number in which the word appears in the sentences. For an example, the index number is 2 for "බල්ලෙකු (a dog)" in S_1 and index number is 1 for "බල්ලෙකු (a dog)" in S_2 . If a particular word is not present in a sentence, we look for similar words using the lexical database. By applying the procedure on S_1 and S_2, the word order vectors (r_1 and r_2) can be

r_1 : { 1 2 3 4 5 }
r_2 : { 2 1 3 4 5 }

obtained:

[5] http://www.languagesdept.gov.lk

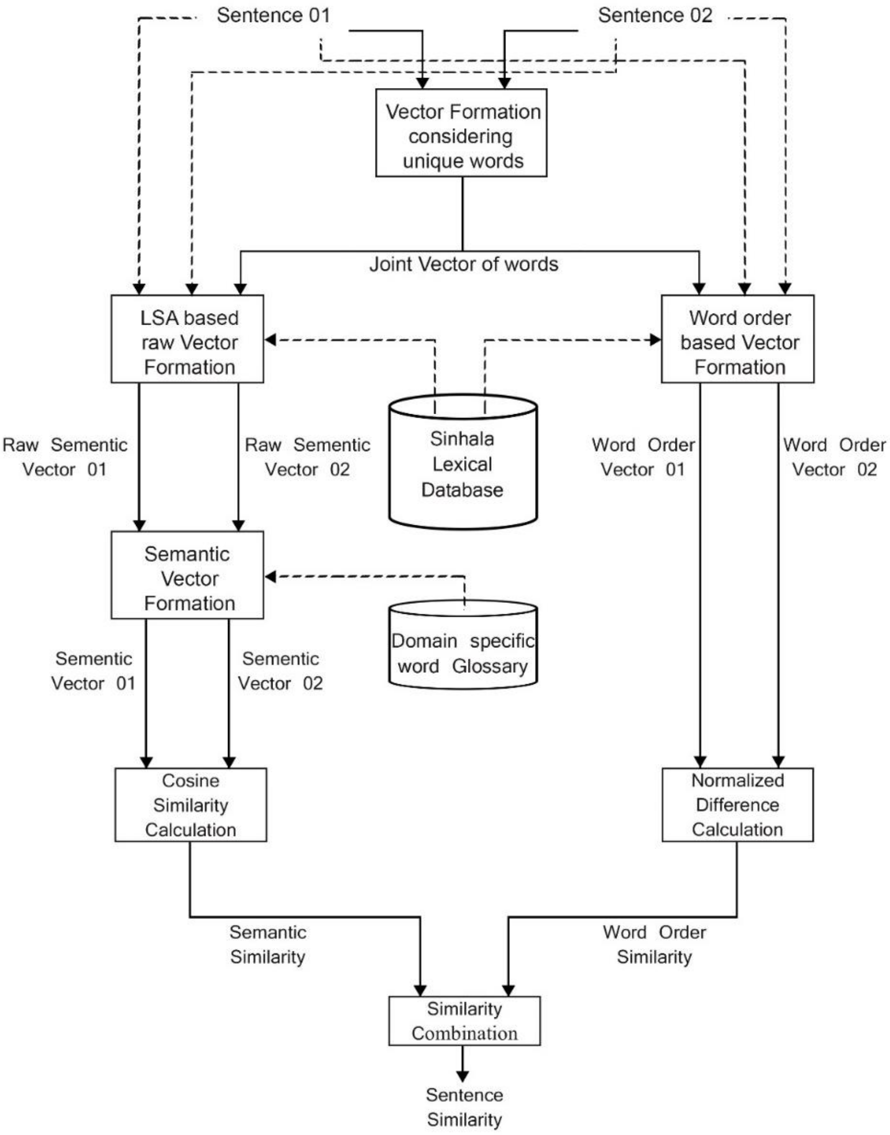

Fig. 1 Overview of the similarity calculation process

Therefore, a word order vector is a basic structure of information of words for a sentence. The task is to measure how similar the word order is. Therefore, we determined the word order similarity (S_r) by the normalized difference of word order as in equation (1). According to Li et. al's (2006), this metric is the best one for indicating the word order in terms of word sequence and location in a sentence.

$$S_r = 1 - \frac{|r_1 - r_2|}{|r_1 + r_2|} \qquad (1)$$

In par with Li et. al's (2006), semantic similarity measure is calculated using corpus-based and knowledge-based similarity measures with the aid of the lexical database and domain specific glossaries, respectively. Relationship between the words is represented by word order based similarity measures. Therefore, combination of these two measures represents both semantic and syntactic information about the short sentences, respectively. Previous researchers have combined many different similarity features using simple weighted average mechanisms (Gomaa, 2012; Mohler, 2011; Li, 2006). Li et. al (2006) combined semantic similarity measures and word order based similarity measures considering only a single weight. Since our approach also requires a single weighted feature combining equation, we adapted Li et. al's (2006) feature combining equation and thus the overall similarity can be calculated as in equations (2) and (3),

$$Sim(S_1, S_2) = T_{val}.S_{1,2} + (1 - T_{val}).S_r \qquad (2)$$

$$Sim(S_1, S_2) = T_{val}.\frac{V_1.V_2}{|V_1|.|V_2|} + (1 - T_{val}).\frac{|r_1 - r_2|}{|r_1 + r_2|} \quad (3)$$

where $T_{val} > 0$ decides the relative contributions of semantic and word order information to the overall similarity computation. Since syntax plays a subordinate role for semantic processing of text, T_{val} should be a value greater than 0.5, i.e. $T_{val} \in (0.5,1]$. We can tune this parameter to any specific domain with minimum effort. For example, when it comes to automatic grading, S_1 would be a student answer sentence and S_2 would be a model answer sentence.

Semantic Similarity Calculation using Word Length Information

When considering the similarity of sentences, word length features also play an important role (Zhao, 2014). For any given two sentences S_1 and S_2, length features record the length information using the following eight measurement functions given in Table 2 as proposed by Zhao (2014). Since these features are language independent, we could directly use them in the context of Sinhala. We created two length vectors (l_1 and l_2) for the sentences S_1 and S_2. Considering these eight length features, we calculated the cosine similarity between the two vectors to form the word length based similarity measures.

Feature	Description
$\boldsymbol{\lvert S_1 \rvert}$	Number of non-repeated words in sentence S_1.
$\boldsymbol{\lvert S_2 \rvert}$	Number of non-repeated words in sentence S_2.
$\boldsymbol{\lvert S_1 - S_2 \rvert}$	Number of unmatched words found in S_1 but not in S_2
$\boldsymbol{\lvert S_2 - S_1 \rvert}$	Number of unmatched words found in S_2 but not in S_1
$\boldsymbol{\lvert S_1 \cup S_2 \rvert}$	Set size of non-repeated words found in either S_1 or S_2
$\boldsymbol{\lvert S_1 \cap S_2 \rvert}$	Set size of shared words found in both S_1 and S_2.
$\boldsymbol{\frac{\lvert S_1 - S_2 \rvert}{S_1}}$	Normalized number of unmatched words found in S_1 but not in S_2
$\boldsymbol{\frac{\lvert S_2 - S_1 \rvert}{S_2}}$	Normalized number of unmatched words found in S_2 but not in S_1

Table 2: Eight length features used in the similarity calculation approach.

Similar to the previous technique, we combined this word length based similarity value ($L_{1,2}$) with the semantic similarity value calculated earlier using a single weight by replacing S_r in equation 2 with $L_{1,2}$. So word length feature based similarity value can be calculated as in equation (4).

$$Sim(S_1, S_2) = T_{val}.\frac{V_1.V_2}{|V_1|.|V_2|} + (1 - T_{val}).\frac{l_1.l_2}{|l_1|.|l_2|} \qquad (4)$$

4 Results and Discussion

Due to space limitations, we only report the results for the hybrid similarity calculation that combined semantic similarity measures with word order based similarity measures, as it gave us the best results.

The hybrid similarity measurement technique discussed in Section 3.3 requires one parameter to be determined before use: the factor T_{val} for weighting the significance between semantic information and syntactic information. Using 1000 sentence pairs, we tuned T_{val} parameter to be 0.87. For the rest of the sentence pairs (4000), we calculated similarity values using our algorithm and compared the results against manually annotated similarity scores. Table 3 shows a comparison of similarities between randomly selected sentence pairs from the 4000 sentence pairs. Even though there are few variations, it can be clearly seen that the two similarity values always represent the same meaning about the sentences and the similarities in Table 3 are fairly consistent with human intuition.

In par with previous research (Bestgen, Biçici, Gupta and Zhao, 2014), we evaluated our results using Pearson (r) and Spearman (ρ) correlation factors along with average Mean Square Error (MSE) for the 4000 sentence pairs. Fig. 2 shows the performance comparison with different values for T_{val}. According to the experimental results, the optimum T_{val} is 0.87 (for English this value is 0.75, for Li et. al's (2006)), results in the lowest average MSE of 0.145. When we compared results reported in previous work done on the SICK dataset (ECNU (Zhao, 2014), CECL ALL (Bestgen, 2014), RTM-DCU (Biçici, 2014), and UoW (Gupta, 2014)), the lowest reported average MSE is 0.325 (Marelli, 2014) whereas our approach gave average MSE of 0.145. We also compared the correlation factors: for the Pearson correlation factor the maximum they could get was 0.828 (Marelli, 2014) whereas our system gave 0.832, and for the Spearman correlation factor they obtained maximum of 0.772 (Marelli, 2014) when our system gave 0.798.

Sentence Pair	**Manually Annotated Score**	**System Generated Score**
A: මිනිසෙකු වාහනයක් අලුත් වැඩියා කරයි (A man is repairing a vehicle) B: මිනිසෙකු මෝටර් රථයක් අලුත් වැඩියා කරයි (A man is repairing a motor car)	0.87	0.75
A: සුනඛයෙක් ඉදිරිය බලාගෙන සිටියි (A dog is looking ahead) B: බල්ලෙකු තණකොළ අතරින් වේගයෙන් දුවයි (A dog is running fast across the grass)	0.23	0.20
A: කුරුල්ලෙකු ජලය මතුපිට සිට පියාසර කිරීමට උත්සාහ කරයි (A bird is trying to fly from the surface of the water) B: පක්ෂියෙක් ගංගාවකට උඩින් පියාසර කරයි (A bird is flying over a river)	0.60	0.53
A: මුවෙක් වැටක් මතින් පනියි (A deer is jumping over a fence) B: මුවෙක් කම්බි වැටක් උඩින් පනියි (A deer is jumping over a wired fence)	1.00	0.85
A: නිල් පැහැති ඇඳුමක් ඇඳ සිටින ටෙනිස් ක්‍රීඩකයා තම ජයග්‍රහණය සමරයි (The tennis player in a blue suit is celebrating his victory) B: ක්‍රීඩකයෙක් පිත්ත ඔසවා ගෙන සතුටින් සිටියි (A player is holding up the bat happily)	0.35	0.29

Table 3: Comparison of similarities between randomly selected sentence pairs

It can be seen that word order similarity calculation has a less impact ((1 - T_{val}) = 0.13) on the final similarity calculation, when compared with English. This is due to the inflection (inflexion) nature of Sinhala. For an example, let's consider the sentence pair S_3 and S_4: the joint word set ($S = S_3 U\ S_4$) for English and Sinhala are {the, man, gives, book, to, child} and {මිනිසා (the man), ළමයාට (to the child), පොත (the book), දෙයි (give), ළමයා (the child), මිනිසාට (to the man)}, respectively. When we form joint vectors for both sentences, it will be exactly similar for the two English sentences, whereas it would be different for the two Sinhala sentences. Here, in English, 'to child' is written as one word 'ළමයාට' in Sinhala, where 'ළමයා' gets inflated into 'ළමයාට' using the dative case.

S_3 : මිනිසා ළමයාට පොත දෙයි (The man gives the book to the child)
S_4 : ළමයා මිනිසාට පොත දෙයි (The child gives the book to child)

The high accuracy of the results may be due to the following reasons: when expressing the same idea, the average word count is high for English than Sinhala due to the high agglutinative behaviour in Sinhala (e.g. "to the honourable president" can be written in one word in Sinhala as "ජනාධිපතිතුමාට"). For the 2500 sentences that we created for Sinhala, the average word count per sentence is 6.694 and for the SICK English dataset used in SemEval 2014, the average word count per sentence is 9.683. Because of this, when we form the semantic vector, we have more information about a single idea using a small number of words. Secondly our lexical resource was created in a way that words similar in meaning are in the same category.

We should also admit that it is not very reasonable to compare the results against that for English, however there is no other way to emphasise our results.

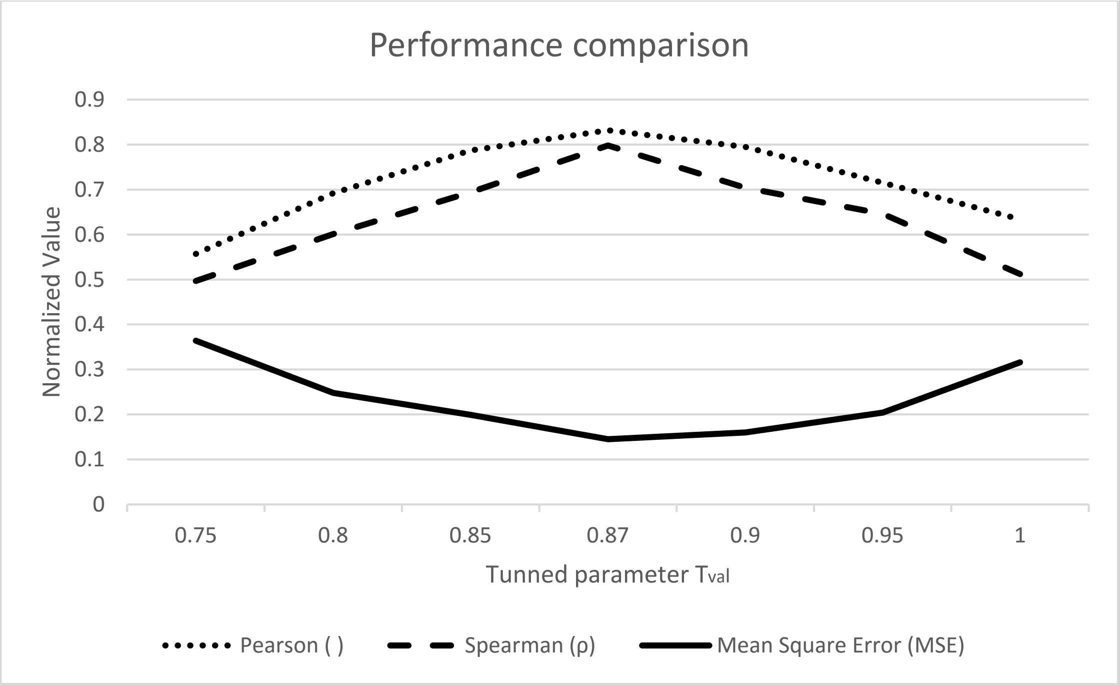

Fig. 2 Graphical representation of performance comparison with different T_{val}

5 Conclusion

We presented the first-ever research on short sentence similarity calculation for Sinhala language. This was carried out using an unsupervised approach based on a hybrid technique, which used semantic similarity measures and word order information. This approach could be implemented because it does not require any complex NLP lexical resources. Therefore, for an under-resourced language such as Sinhala, this is the most suitable way to compare short sentences. Since this technique is largely language independent, the algorithms used for English could be used for Sinhala with only minor modifications.

We found a higher accuracy than what was reported for a comparable dataset for English. Despite the simplicity of the approach used, this result could be partly due to the less average word count in Sinhala short sentences when compared with the same for English short sentences. The best results were given when weight for the word order similarity is 0.13 (1 - T_{val}). Therefore, we can conclude that the word order contribution to short sentence similarity is less for Sinhala, due to the inflection (inflexion) nature of Sinhala.

6 Limitations & Future work

Our lexical database is limited to one to one mappings of similar words, and it does not contain partial similarity values as we have in WordNet. Therefore, our lexical resource should be improved to increase the accuracy of the implemented methodology. Even though our lexical resource consists of multi-words, we do not consider multi-word lookups while creating the semantic vector, which is yet another limitation to be addressed in future research. In order to improve the accuracy furthermore, we plan to test more features for sentence comparison. We also have plans to improve the algorithm to disambiguate word sense using the surrounding words to give contextual information. We also plan to explore different types of short text answers from different domains with varying number of topics in order to prove the generality of our solution.

Acknowledgment

This research is funded by the DL4D 2016 research grant and Prof. V K Samaranayake research grant. The authors would like to thank the typists and annotators for their contribution to the research.

References

Alves, A. O., Ferrugento, A., Lourenço, M., & Rodrigues, F. (2014). Asap: Automatic semantic alignment for phrases. SemEval 2014, 104.

Bestgen, Y. (2014). CECL: a new baseline and a non-compositional approach for the Sick benchmark. Proceedings of SemEval 2014: The 8th International Workshop on Semantic Evaluation (pp. 160-165). Association for Computational Linguistics.

Biçici, E., & Way, A. (2014). RTM-DCU: Referential translation machines for semantic similarity. Proceedings of SemEval 2014: The 8th International Workshop on Semantic Evaluation (pp. 487-496). Association for Computational Linguistics.

Budanitsky, A., & Hirst, G. (2001, June). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources (Vol. 2, pp. 2-2).

Corley, C., & Mihalcea, R. (2005, June). Measuring the semantic similarity of texts. In Proceedings of the ACL workshop on empirical modelling of semantic equivalence and entailment (pp. 13-18). Association for Computational Linguistics.

Gomaa, W. H., & Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. International Journal of Advanced Computer Science and Applications (IJACSA), 3(11).

Gupta, R., Bechara, H., El Maarouf, I., & Orasan, C. (2014, August). UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14) (pp. 785-789).

Hale, M. M. (1998). A comparison of WordNet and Roget's taxonomy for measuring semantic similarity. arXiv preprint cmp-lg/9809003.

Hirst G. and St-Onge D., (1998). Lexical chains as representations of contexts for the detection and correction of malaproprisms, The MIT Press.

Jayasuriya, M., & Weerasinghe, A. R. (2013, December). Learning a stochastic part of speech tagger for sinhala. In Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on (pp. 137-143). IEEE.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics. arXiv preprint cmp-lg/9709008.

Leacock C. and Chodorow M. (1998). Combining local context and WordNet sense similarity for word sense identification. In WordNet, an Electronic Lexical Database, the MIT Press.

Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE transactions on knowledge and data engineering, 18(8), 1138-1150.

Lin D. (1998, July). An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI (Vol. 98, pp. 296-304).

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. SemEval-2014.

Mohler, M., & Mihalcea, R. (2009, March). Text-to-text semantic similarity for automatic short answer grading. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 567-575). Association for Computational Linguistics.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011, June). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 752-762). Association for Computational Linguistics.

Resnik P. (1995). Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. arXiv preprint cmp-lg/9511007.

Tayal, M. A., Raghuwanshi, M. M., & Malik, L. (2014). Word net based Method for Determining Semantic Sentence Similarity through various Word Senses. Proceedings of the First Joint Conference on Lexical and Computational Semantics.

Welgama, V., Herath, D. L., Liyanage, C., Udalamatta, N., Weerasinghe, R., & Jayawardana, T. (2011). Towards a sinhala wordnet. In *Proceedings of the Conference on Human Language Technology for Development*.

Wijesiri, M. Gallage, B. Gunathilaka, M. Lakjeewa, D. Wimalasuriya, G. Dias, R. Paranavithana, N. de Silva (2014). Building a WordNet for Sinhala. In Proceedings of the Seventh Global WordNet Conference, 2014, 100-108.

Wu Z. and Palmer M. (1994). Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico.

Zhao, J., Zhu, T. T., & Lan, M. (2014). Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. Proceedings of the SemEval, 271-277.