Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1874–1881
Marseille, 11–16 May 2020

Word Embedding Evaluation for Sinhala

Dimuthu Lakmal, Surangika Ranathunga, Saman Peramuna, Indu Herath
Department of Computer Science and Engineering, University of Moratuwa
Katubedda 10400, Sri Lanka
(kjtdimuthu.13, surangika)@cse.mrt.ac.lk, (peramunas, induh)@uom.lk

Abstract

This paper presents the first ever comprehensive evaluation of different types of word embeddings for Sinhala language. Three standard word embedding models, namely, Word2Vec (both Skipgram and CBOW), FastText, and Glove are evaluated under two types of evaluation methods: intrinsic evaluation and extrinsic evaluation. Word analogy and word relatedness evaluations were performed in terms of intrinsic evaluation, while sentiment analysis and part-of-speech (POS) tagging were conducted as the extrinsic evaluation tasks. Benchmark datasets used for intrinsic evaluations were carefully crafted considering specific linguistic features of Sinhala. In general, FastText word embeddings with 300 dimensions reported the finest accuracies across all the evaluation tasks, while Glove reported the lowest results.

Keywords: Word Embedding, Sinhala, Evaluation Methodologies

1. Introduction

Distributed representation of words, commonly known as word embeddings, can be considered as one of the preliminary building blocks for most of the downstream Natural Language Processing (NLP) tasks in this era. Word embeddings enable the exploration of fine-grained semantic and syntactic relationships among words by representing each feature as a vector in a low dimensional space (Goldberg, 2017; Elrazzaz et al., 2017). Researchers have been using various distributed representations of words in natural language processing paradigms for decades now. With the rapid growth of neural network based deep learning techniques in recent years, novel neural word embedding approaches have been introduced to overcome the issues of traditional distributional word representations such as Latent Semantic Analysis (LSA). These memory efficient, dense neural word embeddings surpass the performance of traditional sparse word vector representations (Bengio et al., 2003; Mikolov et al., 2013a). Moreover, some researchers obtained word embeddings by combining statistical features used in traditional word embeddings with the features of recent neural word embedding models to improve the performance (Pennington et al., 2014).

NLP community relies on two types of evaluation procedures for word embeddings: intrinsic and extrinsic evaluation. Intrinsic evaluations directly test for syntactic or semantic relationships between words (Schnabel et al., 2015). Intrinsic evaluations are conducted using the datasets crafted based on human assessments on word relations (Bakarov, 2018). Each entry of these datasets consists of related words and a target word or quantified relatedness among words.

On the other hand, extrinsic evaluation tasks measure the performance of word embeddings by using them as inputs to downstream NLP tasks such as sentiment classification, Part-of-Speech (POS) tagging, and Named Entity Recognition (NER) (Pennington et al., 2014; Turian et al., 2010).

Even though there is a growing interest in evaluating different word embeddings based on various linguistic and downstream NLP tasks, only a handful of studies have been conducted to evaluate word embeddings for low-resource languages (Elrazzaz et al., 2017). Furthermore, some of those studies are based on the evaluation datasets that are directly adopted from resource-rich languages such as English (Zahran et al., 2015). Translated evaluation datasets cannot be used to evaluate word embeddings for many languages due to differences in semantic and syntactic relationships between languages. Moreover, word embedding models do not perform in the same manner for every downstream NLP task. Hence, having a proper word embeddings evaluation for a language is important.

Sinhala is a low-resource language that is being used as the most common native language in Sri Lanka, a developing country in South Asia. Sinhala NLP research studies are very much lagging behind compared to advancements in languages such as English. Thus NLP applications such as text clustering (Nanayakkara and Ranathunga, 2018), NER (Manamini et al., 2016), and machine translation (Ranathunga et al., 2018) are still at an investigational level (de Silva, 2019).

On the positive side, some research has already demonstrated the power of word embeddings for end-user NLP tasks such as sentiment analysis (Liyanage, 2019). Here, the use of word embeddings as input features for traditional Machine Learning algorithms such as Logistic Regression, and Deep Learning techniques have given very promising results in the absence of language-specific features such as POS.

This is a major advantage for low-resourced languages such as Sinhala, which do not have resources to develop highly accurate linguistic tools such as POS taggers or Morphological analysers. In particular, to build word embedding models using unsupervised techniques, the only requirement is to have a sufficiently large monolingual corpus, which is not so much of a demanding need for many languages including Sinhala in this era. On the negative side, building and evaluating word embedding models is a time-consuming task, where the

models have to be evaluated in aspects such as vector dimensionality.

So far, for Sinhala, some Word2Vec and FastText models are available, which have been developed based on different data sources [1]. However, no systematic evaluation of these word embedding models similar to what has been done for languages such as English (Ghannay et al., 2016) and Arabic (Elrazzaz et al., 2017), has been done. These studies cannot be directly applied for Sinhala, due to the vast differences between the languages.

This paper presents the first systematic evaluation of word-based word embedding models for Sinhala. First, we prepared a clean corpus using a subset of the Common Crawl dataset[2], and then performed two intrinsic evaluation tasks (word analogy and relatedness) and two extrinsic evaluation tasks (sentiment analysis and POS tagging) on three types of word embeddings: Word2Vec, Glove, and FastText, which were trained on the cleaned Common Crawl corpus. Benchmark datasets for intrinsic evaluations are crafted by two linguists and two native speakers. In general, the results are the highest in FastText, and it decreases in the order of Word2Vec and Glove. The cleaned common crawl corpus, word embedding models, as well as the benchmark datasets have been publicly released[3].

The rest of the paper is organized as follows. Section 2 reports related work, and Section 3 presents the experimental setup. Sections 4 and 5 report intrinsic and extrinsic experiments, respectively. Section 6 discusses the achieved results, and finally Section 7 concludes the paper.

2. Related Work

2.1. Word Embeddings

Advancements of neural networks have led researchers to explore a variety of approaches for deriving word embeddings. In this study, we evaluate three types of most widely used unsupervised word embeddings in the NLP community: Word2Vec, Glove and FastText. All these word embedding models have been successfully used in a wide range of NLP tasks in English (Lai et al., 2015).

2.1.1. Word2Vec

Word2Vec (Mikolov et al., 2013a) is a neural word embedding model trained on a simple feed forward neural network. It has heavily contributed towards the recent success of many NLP applications because of its simple structure, reduced complexity, and the ability to build dense low-dimensional vector representations (Mikolov et al., 2013b). Mikolov et al. (2013a) introduced two types of word embeddings namely: continuous bag of words (CBOW) model and continuous skip-gram model. Even though both these methods ignore the word order information, they have shown competencies of capturing the semantic and syntactic relations among words.

2.1.2. FastText

When deriving word embeddings, it is important to consider morphology of words, particularly for morphologically rich languages such as Sinhala. Addressing this issue, FastText was introduced as an extension of skip-gram Word2Vec model, where it represents words as a bag of character n-grams instead of the word as a whole (Bojanowski et al., 2017). In FastText, vector representation of a word is derived as the vector summation of character n-grams. Thus, a vector representation can be obtained even for rare words that are not exposed at training time of the FastText model.

2.1.3. Glove

While both Word2Vec and FastText are predictive neural models, Glove is a count-based model that also exploits some key features in those predictive models. Pennington et al. (2014) revealed Glove as a new global log bilinear regression model that combines advantages of two major embedding methods: global matrix factorization and local context window. The intuition behind Glove is that the ratios of co-occurrence probabilities among words have the potential of encoding some kind of a relation among words.

2.2. Evaluating Word Embeddings

Previous research on word embeddings evaluation has suggested a variety of evaluation methodologies. These can be broadly categorized as intrinsic and extrinsic evaluations.

It is worth noting that extrinsic or intrinsic evaluation alone cannot guarantee the quality of word embeddings. Different NLP applications have substantial differences between each other in various aspects, and the usage of word embedding features (Luong et al., 2015; Lopez and Kalita, 2017). On the other hand, performance indications given only based on intrinsic evaluations do not necessarily explain how well word embedding models perform for downstream NLP tasks. Thus, performance of one NLP task may not reflect the performance of other NLP tasks, and it is essential to preform both extrinsic and intrinsic evaluations in order to present a proper word embedding evaluation benchmark.

2.2.1. Intrinsic Evaluation

Intrinsic evaluation methods can be categorized into four broad categories: Relatedness, Analogy, Categorization and Selectional preference (Schnabel et al., 2015). Since relatedness and analogy tasks are the most widespread evaluation methods, we focused on those two methods in this research.

Analogy tasks consist of sets of semantic and syntactic relationship questions. Each question includes a set of query words and a target word. A question is correctly answered only if the proposed word by the embedding model is similar to the target word. The most popular analogy dataset is released by Mikolov

[1] https://fasttext.cc/docs/en/crawl-vectors.html

[2] https://commoncrawl.org/

[3] https://github.com/nlpc-uom/WEIntrinsicEvaluation

et al. (2013a), which consists of five types of semantic questions and nine types of syntactic questions. BATS (Bigger Analogy Test Set) is another comprehensive analogy dataset (Gladkova et al., 2016) with a balanced set of analogy questions (99,200 questions in total) across four categories of relations: inflection and derivational morphology, and lexicographic and encyclopedic semantics. Authors have avoided homonyms and ambiguous words, and extracted queries from various existing datasets in order to make the dataset balanced and consistent.

Relatedness tasks measure to which degree the word embedding model captures any kind of semantic relation between word pairs. Datasets used for this task consist of sets of word pairs and their relatedness scores assigned by human evaluators. The most widely used data sets for this task are WordSim-353 (Finkelstein et al., 2002), and MEN (Bruni et al., 2014) that contain 353 word pairs and 3000 word pairs, respectively. In this task, embedding models rate the semantic proximity of two words in terms of cosine similarity metric, and measure the correlation (Spearman or Pearson) with human relatedness value (Schnabel et al., 2015).

2.2.2. Extrinsic Evaluation

Word embeddings could be used in almost any downstream NLP application. Extrinsic evaluation tasks measure the ability of using word embeddings in those downstream NLP tasks (Bakarov, 2018). The performance of an NLP task is considered as a proxy for the performance of word embeddings. Frequently used NLP tasks for evaluations are POS tagging, noun phrase chunking, NER, and text classification (Pennington et al., 2014; Turian et al., 2010).

3. Experimental Setup

3.1. Corpus

Since Sinhala is a low-resourced language, it is challenging to prepare a large cleaned corpus that is comparable with English for any experiment conducted in the domain of Sinhala NLP. With the initiative of Common Crawl (CC), the NLP community now has the access to petabytes of multilingual data collected over 8 years of web crawling. Common Crawl can be considered as a precious starting point for building a cleaned large corpus for Sinhala. Common Crawl monthly dataset only contains 0.007% of content in Sinhala[4], however, this amount is still significant compared to other publicly available Sinhala corpus datasets.

A pre-processed version of the CC is available, where the web related tags have been removed. We further preprocessed this corpus to remove characters that are not related to Sinhala language, and to remove punctuation and special characters. We prepared two corpora: with and without stop words. Two linguists in Sinhala contributed to preparing the stop word list. The final pre-processed corpus with stop words consists of 94,648,911 tokens. There are 11,114,600 of stop word occurrences in the corpus.

[4]https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

3.2. Building Word Embedding Models

We measure the performance of three types of word embedding models: Word2Vec, FastText and Glove, based on two intrinsic and two extrinsic evaluation tasks. Both CBOW and skip-gram models of Word2Vec are evaluated in this study. FastText model was trained based on the Skip-gram architecture. All the word embedding models were trained on the cleaned CC dataset. In addition, we evaluated the pre-trained FastText model (based on the CBOW architecture) released by Facebook(Inc, 2020), which was trained on CC and Wikipedia data. Hyperparameters of the trained word embeddings and the pre-trained Facebook FastText model are summarized in Table 1. Intrinsic and extrinsic evaluations were conducted using the vector dimensions of 100, 200 and 300 for the corpus that contains stop words. According to the results, the best accuracy is given when we use embeddings with 300 dimensions. Therefore, we ran intrinsic and extrinsic evaluations for the corpus without stop words, only using the embeddings with 300 dimensions. Results obtained for the analogy task evaluation are shown in Table 3. Results of the FastText model released by FaceBook are also given.

4. Intrinsic Evaluation

Analogy and related tasks were conducted as intrinsic evaluations.

4.1. Intrinsic Evaluation using Analogy

The analogy questionnaire used in this study is inspired by the BATS dataset. Most of the translated BATS questions cannot be directly applied for evaluating Sinhala word embeddings due to several reasons. First, a significant number of words in the BATS dataset are rare or non-existent words in Sinhala, and it is most likely that those words do not exist in the corpus. For instance, some words that are used to represent sounds of animals such as 'bray', 'bleat' and 'oink', do not have corresponding words in Sinhala. Moreover, most of the derivational morphologies included in the BATS dataset cannot be observed in Sinhala (e.g. : first(noun+less), fourth (over+adj./verb) and sixth (re+verb) derivational relationships in BATS dataset).

Still there are plenty of other derivational morphological relationships that can be observed in Sinhala, due to its morphological richness. Thus, we replaced most of the original derivational relations with new relations that are applicable to Sinhala. Most of the verbs included in the BATS dataset cannot be expressed as a single word in Sinhala. In general, verbs are represented as compound verbs consisting of two or more words. For instance, 'believe' and 'develop' are represented as `විශ්වාස කරනවා' (vishvāsa karanavā) and

Embeddings	Dimension	Window	Neg.	Min. Cnt.	Dwn. Sam.	Max count,alpha
Word2Vec CBOW	100,200,300	9	5	1	0.001	-
Word2Vec Skip-gram	100,200,300	9	5	1	0.001	-
FastText	100,200,300	9	5	1	0.001	-
Glove	100,200,300	10	-	-	-	100, 0.75
Pretrained FastText	300	5	10	N/A	N/A	-

Table 1: Hyper Parameters of Word Embedding Models

[1] Neg. = Negative Sampling, Min. Cnt = Minimum count, Dwn. Sam = Down Sampling.

`දියුණු කරනවා' (diyunu karanavā), respectively in Sinhala. Since we only focus on embeddings for single words in this study, most of the original verbs in the BATS dataset were replaced with the verbs that can be represented only using a single word (e.g.: `ගයනවා' (gayanavā), which corresponds to the verb 'sing').

Relationships that belong to lexicographic semantics were avoided, since the lexicographic semantics concept is related to word relatedness. We evaluate the performance of word embeddings in terms of capturing the degree of relatedness among words as a separate intrinsic task, which will be further described below. Relationships of the newly developed Sinhala analogy dataset are shown in Table 2.

The new analogy questionnaire prepared for Sinhala consists of 8 inflectional and 8 derivational morphological relationships, and 6 encyclopedic semantic relationships. Each analogy question consists of two word pairs: (W1, W2) and (W3, W4), where the two word pairs are taken from the same relation. Every possible combination of word pairs from the same relation is generated and prepared as a analogy question. Thus, 27,382 analogy questions were prepared from inflectional and derivational categories by combining word pairs from the same relation. Further, encyclopedic category consists of 5,364 analogy questions. To answer an analogy question, the given word embedding model calculates a target word vector t, using word vectors, $w1$,$w2$ and $w3$, corresponding to the remaining three words, W1,W2 and W3 in the given analogy question, as shown in equation 1. Then the system retrieves the word that has the closest cosine similarity with respect to the target word. If the retrieved word is similar to the word W4, then the answer given by the embedding model is considered correct. Further, if at least one word in the question does not exist in the embedding model, we consider the question is wrongly answered.

$$t = w2 - w1 + w3 \quad (1)$$

We altered the procedure of originally proposed analogy evaluation by Mikolov et al. (2013a) by taking the top 5 answers to a given target word into consideration. If the target word is among those top 5 words, then the answer given by the embedding model is considered correct. Making the evaluation criteria less strict is particularly important for morphology-rich languages such as Arabic and Sinhala for multiple reasons (Elrazzaz et al., 2017).

In Sinhala, rather than representing various word forms by phrases or a completely different word, different forms of a word are derived by adding prefixes or suffixes. For example, 'to king' and 'to queen' are represented as `රජතුමාට' (rajathumāta) and `බිසවට' (bisavata) by adding `ට' (ta) suffix to words `රජතුමා' (rajathumā) and `බිසව' (bisava) which stand for 'king' and 'queen' in Sinhala. Most importantly, there are multiple ways of deriving word forms that give the same meaning by adding various suffixes and prefixes. (e.g. even though, the only way of representing plural word of 'man' is 'men', there are multiple ways of representing the plural word of 'man' (`මිනිසා' (minisā)) in Sinhala, as `මිනිසුන්' (minisun) and `මිනිස්සු' (minissu)).

Results obtained for the analogy task evaluation is shown in Table 3. Results of the FastText model released by FaceBook are also given.

4.2. Relatedness Intrinsic Evaluation

The relatedness questionnaire is prepared based on the widely used WordSim353 data set that consists of 353 word pairs (Finkelstein et al., 2002), and their semantic relatedness scores assigned by human annotators. Since most of the words in the original dataset are not affected by the constraints described in Section 4.1., 87% of the word pairs are directly translated to Sinhala in order to construct the Sinhala relatedness questionnaire. The new dataset includes 345 word pairs. We did not rely on exact scores assigned in the original word pairs since the scores depend on the language and the usage of words in the society. For example, the word 'president' `ජනාධිපති' (janādípathi) and 'medal'`පදක්කම' (padakkama) are more related in the context of Sri Lanka. President medals are awarded to various individuals frequently in Sri Lanka, so that it is being reported in newspaper articles very often. Two Sinhala native speakers were involved in an independent annotation of relatedness scores for each word pair. Final score for each word pair is calculated by averaging the scores assigned by the two annotators.

Cosine similarities were calculated between the two words of a word pair in the relatedness dataset using trained word embedding models. Next, the Spearman correlation was calculated between the cosine similarities and scores given by human annotators. Table 4 reports results we obtained from each word embedding.

Category	Relationship	Example
Inflections	I01: Plural	ඇල්බමය, ඇල්බම් (ælbamaya, ælbam)
	I02: Superlative	නිවැරදි, නිවැරදිම (niværadi, niværadima)
	I03: Present Plural: Present Singular	පිළිගනිති, පිළිගනියි (pi□iganiti, pi□iganiyi)
	I04: Present Plural: Present Participle	උගන්වති ,උගන්වමින් (uganvati, uganvamin)
	I05: Present Plural: Past Singular	සපයති, සැපයුවේය (sapayati,sæpayuv□ya)
	I06: Participle: Present Singular	කරමින්,කරයි (karamin, karayi)
	I07: Participle: Past Singular	ලබමින්,ලැබුණේය (labamin, læbu□□ya)
	I08: Present Singular: Past Singular	දෙයි,දුන්නේය (deyi, dunn□ya)
Derivational	D01: Adjective: Antonym Adjective	ලස්සන, අවලස්සන (lassana, avalassana)
	D02: Adjective: Adverb	ශක්තිමත්,ශක්තිමත්ව (□aktimat, □aktimatva)
	D03: Past Participle: Negation Past Participle	ලියූ,නොලියූ (liy□,noliy□)
	D04: Adjective: Noun	අස්ථාවර,අස්ථාවරත්වය (asth□vara, asth□varatvaya)
	D05: Verb: Noun	මවයි,මැවීම (mavayi,mæv□ma)
	D06: Verb: Gerund	සපයති,සපයන්නා (sapayati, sapayann□)
	D07: Verb: Verbal Noun	ගයති,ගැයුම (gayati, gæyuma)
	D08: Noun: Dative	වනය,වනයට (vanaya, vanaya□a)
Encyclopedic	E01: Country: Capital	බර්ලින්,ජර්මනි (barlin, jarmani)
	E02: Name: Nationalities	ඇරිස්ටෝටල්,ග්‍රීක (æris□□□al,gr□ka)
	E03: Things: Colors	ඇපල්,රතු (æpal,ratu)
	E04: Male: Female	නළුවා,නිළිය (na□uv□,ni□iya)
	E05: Province: Capital	දකුණ,ගාල්ල (daku□a,g□lla)
	E06: Institute: Head	පාසල,විදුහල්පති (p□sala,viduhalpati)

Table 2: Relationships and Examples of Sinhala Analogy Dataset

Rel.	With Stop Words												Without Stop Words				Pre. FT
	CBOW			Skip-gram			Glove			FastText			CB	Skip	Gl	FT	
	100	200	300	100	200	300	100	200	300	100	200	300	300	300	300	300	300
I01	33.5	43.0	**43.6**	25.3	28.8	27.1	12.0	6.07	7.15	25.3	28.6	27.1	40.3	26.0	6.55	26.0	20.9
I02	28.8	37.3	40.1	32.7	31.8	61.9	10.13	8.4	8.2	45.6	58.8	**62.2**	38.2	23.3	8.07	61.2	50.6
I03	18.7	24.0	24.8	17.4	20.9	26.7	8.4	8.73	10.6	18.2	27.0	**27.8**	24.7	17.0	9.84	26.0	9.2
I04	28.9	37.4	39.2	23.9	26.0	47.8	12.93	7.53	9.86	37.6	**50.0**	48.8	37.2	20.7	8.7	**50.0**	26.7
I05	13.0	16.3	24.8	9.3	10.5	29.6	3.87	4.07	4.52	17.5	28.7	**30.3**	18.7	9.4	4.22	27.0	14.6
I06	27.2	34.0	**39.2**	22.9	24.6	24.4	11.53	7.07	6.6	19.2	24.3	25.0	36.8	21.8	6.73	23.9	14.5
I07	17.7	20.7	17.2	14.4	15.7	36.4	6.68	4.93	5.54	26.3	**36.9**	36.5	23.5	13.5	5.06	33.0	26.3
I08	9.4	12.9	**35.1**	10.1	10.8	29.6	8.8	8.53	9.18	18.8	26.5	30.1	14.6	9.1	8.57	28.7	17.0
All	21.8	27.8	28.9	18.9	20.7	34.1	9.3	6.92	7.59	25.0	33.8	**34.7**	28.6	17.3	7.11	33.3	20.9
D01	7.3	14.5	15.2	11.0	16.0	21.9	10.93	8.87	8.2	11.2	19.0	21.9	16.0	13.8	7.99	**22.8**	21.3
D02	18.2	25.8	29.0	22.4	23.6	46.7	5.65	3.93	4.58	36.9	**50.1**	49.0	31.5	19.6	3.9	43.6	45.1
D03	7.3	8.7	8.9	5.2	6.5	31.8	3.27	2.13	2.75	12.4	24.2	**34.0**	9.4	5.3	2.61	28.4	20.9
D04	9.0	12.7	13.1	6.6	4.5	35.4	0.4	0.1	0.2	28.3	35.8	**35.9**	9.8	1.6	0.21	31.5	23.1
D05	17.7	31.0	**32.2**	11.3	14.6	21.7	5.87	3.13	3.97	10.4	21.6	21.7	25.0	11.1	3.64	21.3	4.4
D06:	7.3	12.7	14.6	1.7	1.0	15.2	1.72	1.83	1.85	7.5	14.1	15.5	4.5	0.9	1.83	8.2	**18.6**
D07:	1.3	1.6	**1.8**	0.8	1.0	0.6	0.53	0.53	0.52	0.45	0.53	0.8	0.9	0.6	0.49	1.0	0.6
D08	67.3	71.2	72.0	56.8	61.7	64.3	23.55	15.59	20.54	53.7	67.5	64.3	**72.2**	54.7	19.11	69.9	51.9
All	15.3	21.4	22.3	13.0	15.0	28.0	6.36	4.44	5.84	17.7	26.9	**28.7**	20.1	12.4	5.6	26.8	20.6
E01	8.8	11.7	14.0	14.2	16.7	16.3	3.33	2.27	3.09	13.2	17.8	**16.8**	14.8	14.4	2.81	17.6	6.0
E02	13.1	24.0	26.7	13.0	21.4	10.4	16.59	12.62	12.06	4.2	11.8	11.0	**27.5**	17.5	11.77	9.6	4.1
E03	1.9	4.0	4.74	8.8	**13.7**	8.17	21.98	17.84	20.6	2.52	7.7	8.17	3.3	10.1	19.14	7.8	2.2
E04	28.3	32.5	**34.6**	26.5	26.9	29.3	20.63	15.24	16.93	27.2	31.4	29.8	33.0	24.8	17.77	29.4	14.13
E05	39.3	66.1	62.5	32.1	**69.6**	50.0	14.29	7.14	8.14	23.2	44.6	50.0	62.5	57.1	7.55	50.0	12.5
E06	16.0	26.9	**31.4**	21.8	21.8	10.3	5.13	3.12	5.15	12.8	16.0	10.3	27.6	11.5	5.07	8.3	6.4
All	13.6	19.1	**21.1**	16.2	20.3	16.7	14.41	10.93	12.7	12.5	18.0	17.1	20.8	17.1	13.14	**16.7**	6.8

Table 3: Results of Analogy Task. (Accuracies in Percentage)

[1] Values in 'All' rows represent average acccuracies for Inflectional, Derivational and Encyclopedic relation categories. CB = CBOW, Skip = Skip-gram, Gl = Glove, FT = FastText, Pre. FT = Pretrained FastText

[2] Best result for each relationship is indicated in bold. In general, Word2Vec CBOW and FastText with 300 dimensions show better results.

5. Extrinsic Evaluation

As previously explained, still a very limited number of downstream tasks have been carried out in Sinhala NLP domain successfully due to resource sparsity. Thus, we could only perform two extrinsic evaluation task for Sinhala word embedding models.

First evaluation was conducted by following a previous sentiment analysis assessment[5]. In particular, it presents a sentiment analysis system for Sinhala news comments to classify each comment into two categories: positive, and negative. Further, it presents a rigorous analysis of the sentiment analysis performance with regards to different word embedding types and machine learning techniques. The annotated news comments dataset [6] contains comments from 276 news articles representing a wide variety of news categories including politics, sports, crime, economy and culture. Authors have experimented with both statistical machine learning algorithms and Deep Neural Network models. Best results have been obtained from logistic regression with regards to statistical supervised models, while an RNN-LSTM model reported the best

[5] https://github.com/suralk/Thesis

[6] https://github.com/theisuru/sentiment-tagger

	With Stop Words			Without S.W
Dimen.	**100**	**200**	**300**	**300**
CBOW	0.528	0.546	0.548	0.573
Skipgram	0.625	0.635	0.640	0.643
Glove	0.389	0.402	0.415	0.428
FastText	0.597	0.641	0.644	**0.650**
Pre. FT	0.611			

Table 4: Results of Relatedness Task (Spearman Correlations

[1] Pre. FT = Pretrained FastText

	With Stop Words			Without S.W
Dimen.	**100**	**200**	**300**	**300**
CBOW	84.22	83.83	83.94	85.33
Skipgram	87.26	86.71	87.34	87.21
Glove	82.96	84.16	84.34	82.91
FastText	86.40	86.58	**87.76**	87.48
Pre. FT	84.98			

Table 5: Results of the Sentiment Analysis Task (F1 Scores %)

[1] Pre. FT = Pretrained FastText by FaceBook

overall results out of all statistical and neural network based models. Hence, the RNN-LSTM model that gave the best results was used to build our word embedding benchmark. Results obtained by running the sentiment analysis is reported in Table 5.

Secondly, we experimented with a neural network based POS tagger. We used the existing implementation[7] of a combined approach of bidirectional Long Short-Term Memoey (LSTM) network and Conditional Random Field (CRF) was trained using a labeled Sinhala POS tag dataset (Fernando and Ranathunga, 2018; Fernando et al., 2016). Bidirectional LSTMs have been proven as a successful sequence modeling architecture in recent years (Plank et al., 2016). CRF is a probabilistic approach for sequence modelling, which was widely used before Deep Learning models attracted the attention of NLP community. The proposed network can efficiently utilize both past and future input information via biLSTM layers, and sentence level tag information via a CRF layer. The training dataset consists of 28,630 sentences and testing dataset consists of 3,182 sentences. Training and test set remain unchanged for every tested word embedding model. The obtained results are shown in Table 6.

6. Discussion

Not having a proper word embedding evaluation for Sinhala has made selecting appropriate word embedding model for various NLP tasks is challenging. Thus, we implemented the first embedding evaluation benchmark and evaluated three types of word embeddings

[7] https://github.com/wantinghuang/tensorflow-lstmcrf-postagger

	With Stop Words			Without S.W
Dimen.	**100**	**200**	**300**	**300**
CBOW	86.30	88.60	89.64	87.68
Skipgram	84.94	87.18	87.44	86.84
Glove	83.02	85.91	87.10	85.29
FastText	88.78	90.85	**90.97**	89.33
Pre. FT	88.48			

Table 6: Results of the POS Tagging task (Accuracy %)

[1] Pre. FT = Pretrained FastText by FaceBook

trained on a Sinhala corpus. Evaluation criteria is designed for both intrinsic and extrinsic evaluations based on the datasets prepared by Sinhala native speakers and linguists. Analogy and relatedness tasks, were conducted as intrinsic evaluation while sentiment analysis and POS tagging were conducted as extrinsic evaluation.

FastText models trained on Common Crawl data reported the best overall accuracy for inflectional and derivational relationships, while CBOW reported the best overall accuracy for Encyclopedic relationships. All embedding types have not captured many analogy relations, particularly derivational and encyclopedic relations. More specifically, verb to noun conversion relationships (D06 and D07) can be observed as difficult relationship categories for all types of word embeddings. Relationships that have poor and good results are mostly similar across all the embedding types. It indicates conceptual similarities that exist among word embeddings.

Similar to the results reported by many of the previous studies, skip-gram approach yields better results for one downstream NLP task and one intrinsic task compared to the CBOW approach, as CBOW assigns limited probabilities to rare words at the training phase of the model. However, even though skip-gram results are higher in relatedness evaluation and sentiment analysis task, the difference of the performances drops for the corpus without stop words, compared to the corpus consisting of stop words. Moreover, performance of the CBOW model on the analogy evaluation and POS tagging task is more superior than that of the skip-gram model. According to this observation, we can assume that the perfect word embedding for a given NLP task can be varied according to nature of the task. Furthermore, above results also explain that having better performance for intrinsic tasks does not necessary prove that it performs well in extrinsic tasks and vise versa.

Intuitively, representing word vectors by more dimensions leads to capture more linguistic features among words. The highest performance was reported by the models with 300 vector dimensions except in few scenarios of the analogy task.

FastText trained on our preprocessed CC has significantly outperformed the pre-trained Facebook FastText model. This can be attributed to multiple reasons. Our FastText model is trained using a context

window of 9 while Facebook FastText model trained on a context window of 5. As explained earlier, verbs and nouns are expressed as phrases of two or more words in Sinhala. Thus, commonly used context widow of 5 apparently is not enough to represent a semantically meaningful context. Moreover, Facebook FastText is trained based on the CBOW architecture, which as previously discussed, is not the ideal architecture in scenarios when resources are limited. In addition, the preprocessing steps applied to the CC dataset possibly had a positive contribution towards the higher accuracy of our FastText model over the pretrained FastText model.

Removing stop words could not gain a performance boost over the analogy task and sentiment analysis task. Surprisingly, the word embedding models based on the corpus without stop words yield best results for the relatedness task.

In general, our intrinsic and extrinsic evaluations provide evidence to prove that FastText trained on CC is the best word embedding for Sinhala compared to Word2Vec and Glove. This may be mainly due to the fact that FastText takes sub word information into consideration when generating word vectors. Sinhala is a highly inflectional language. For instance, there are five ways of adding suffixes to generate the past tense of a verb according to the subject type of the sentence. For example, past tense of the verb 'go' can be expressed as either of `ගියේය' (giyēya), `ගියෝය'' (giyōya), `ගියහ' (giyaha), `ගියාය' (giyāya) and `ගියෙමි' (giyemi), based on the subject type of the sentence. Results indicate that FastText has successfully exploited this sub word information in Sinhala. In general, Glove reported the lowest results for all the intrinsic and extrinsic tasks. This can be attributed to the size of the corpus we used in this study, despite the fact that we leveraged one of the largest Sinhala corpus for this experiment.

7. Conclusion

In this research, we carried out a rigorous analysis of word embedding models: Word2Vec, Glove and FastText for Sinhala. This study is the first word embedding evaluation benchmark for Sinhala. FastText word embedding with 300 vector dimensions reported the overall best results, hence proving the importance of sub-word information in morphologically rich languages like Sinhala. Moreover, it is clear that there is no universal word embedding type that gives the finest performance for every NLP task. Having better results for intrinsic tasks does not guarantee obtaining better results for different NLP tasks. Further, removing stop words from the corpus is highly subjective on the task that we are going to apply word embeddings to.

The benchmark described in this paper can be further enhanced by evaluating more novel word embedding types such as BERT (Devlin et al., 2018) and XLnet (Yang et al., 2019). Evaluating phrase embeddings would be a crucial next step for Sinhala, since most of the verbs in Sinhala include two or more words. Moreover, introducing and evaluating word embeddings that exploit morphological and syntactic features such as POS tags will further extend the discussion on Sinhala word embeddings.

8. Acknowledgment

Authors would like to thank the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education, Sri Lanka funded by the World Bank, for supporting this project.

9. References

Bakarov, A. (2018). A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

de Silva, N. (2019). Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elrazzaz, M., Elbassuoni, S., Shaban, K., and Helwe, C. (2017). Methodical evaluation of arabic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–458.

Fernando, S. and Ranathunga, S. (2018). Evaluation of different classifiers for sinhala pos tagging. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 96–101. IEEE.

Fernando, S., Ranathunga, S., Jayasena, S., and Dias, G. (2016). Comprehensive part-of-speech tag set and svm based pos tagger for sinhala. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 173–182.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Ghannay, S., Favre, B., Esteve, Y., and Camelin, N. (2016). Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Inc, F. (2020). Word vectors for 157 languages. `https://fasttext.cc/docs/en/crawl-vectors.html`. Accessed: 2020-03-03.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Liyanage, I. (2019). Sentiment analysis of sinhala news comments. `https://github.com/suralk/Thesis`. Accessed: 2020-03-03.

Lopez, M. M. and Kalita, J. (2017). Deep learning applied to nlp. *arXiv preprint arXiv:1703.03091*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Manamini, S., Ahamed, A., Rajapakshe, R., Reemal, G., Jayasena, S., Dias, G., and Ranathunga, S. (2016). Ananya-a named-entity-recognition (ner) system for sinhala language. In *2016 Moratuwa Engineering Research Conference (MERCon)*, pages 30–35. IEEE.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nanayakkara, P. and Ranathunga, S. (2018). Clustering sinhala news articles using corpus-based similarity measures. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 437–442. IEEE.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. page 412.

Ranathunga, S., Farhath, F., Thayasivam, U., Jayasena, S., and Dias, G. (2018). Si-ta: Machine translation of sinhala and tamil official documents. In *2018 National Information Technology Conference (NITC)*, pages 1–6. IEEE.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. pages 5754–5764.

Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., and Atyia, A. (2015). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–443. Springer.