Hybrid approach for semantic similarity calculation between Tamil words

Deepa Karuppaiah* and P.M. Durai Raj Vincent

Vellore Institute of Technology,
Vellore, Tamilnadu, India
Email: deepa.k@vit.ac.in
Email: pmvincent@vit.ac.in
*Corresponding author

Abstract: Semantic similarity, sometimes referred as semantic relatedness, is one of the important concepts that help in various applications that involve natural language processing. In literature, there are plenty of similarity measures to compute the relationship among words in monolingual and cross-lingual documents. They help us in understanding text, finding plagiarism, information retrieval etc. They can be categorised based on the resources used into corpus-based and knowledge-based measures. These measures are plenty for the English language. For the Tamil language, there are hardly any works in calculating the similarity between words. In this paper, we proposed a similarity finding technique that exploits the knowledge from the resources like Tamil Indo WordNet, Tamil Wikitionary and Oxford Tamil Dictionary. We have used the definitions and example sentences of each word that are available through each of these resources for similarity calculation. The proposed approach is evaluated using human evaluated Miller Charles and Rubenstein Goodenough datasets.

Keywords: semantic similarity; Tamil words similarity; Indo WordNet; knowledge-based similarity.

Reference to this paper should be made as follows: Karuppaiah, D. and Vincent, P.M.D.R. (2021) 'Hybrid approach for semantic similarity calculation between Tamil words', *Int. J. Innovative Computing and Applications*, Vol. 12, No. 1, pp.13–23.

Biographical notes: Deepa Karuppaiah received her BE and ME from Anna University, Chennai, India. She is presently working as an Assistant Professor (Senior Grade) in the School of Information Technology and Engineering at Vellore Institute of Technology (VIT), India. She is having more than 15 years of teaching experience. Her research interest includes machine learning, information retrieval and operating systems.

P.M. Durai Raj Vincent received his BE and ME from Anna University, Chennai, India. He also received his PhD from VIT University Vellore. He is presently working as an Associate Professor in the School of Information Technology and Engineering at Vellore Institute of Technology (VIT), India. He is having more than 13 years of teaching experience. His current research interest includes security, machine learning and data analytics.

1 Introduction

The ability to measure the relatedness between two words in any language is one of the most important sub-domain and the most promising support in entire natural language processing (NLP) research. To understand any sentence semantically, we would need to understand the contextual words and their relationships in NLP. This metric that assess the common things between two words supports the other related areas like information retrieval, and knowledge discovery. To understand any sentence, we would employ different statistical models like n-gram or ANN models (Mezzoudj and Benyettou, 2018). The relatedness or similarity score between two words will help us in understanding the importance of those words in a context and the context itself. In reality, usually one concept would be represented using many different words in a language. When comes to application of these words, we may be restricted to use all of these words with respect to the context in which it has to be applied. Hence, we need to find the semantic relatedness of one word with the other. And this similarity can solve so many problems in NLP. Various similarity measures have been proposed and applied in domains including information retrieval [product information retrieval (Akmal et al., 2014), collaborative tagging (Uddin et al., 2013)], query expansion (Azad and Deepak, 2019; Nasir et al., 2019), medical term disambiguation (Gabsi et al., 2017), software development (Alenazi et al., 2018), e-mails suggestion (Pera and Ng, 2007) and plagiarism detection [plagiarism in English documents (Ehsan and Shakery, 2016), Malayalam documents (Sindhu and Idicula, 2017)]. This field of study is one of the well analysed areas especially for English language. Though there are many researchers have studied the similarity problem of two words in English, Tamil has no such study to find the semantic similarity between words. In this paper, we proposed a hybrid approach that

deploys various Tamil dictionaries and Indo WordNet (Bhattacharyya, 2010) Tamil to calculate the similarity score between two Tamil words.

Based on the sources used to compute the similarity, the similarity measures can be categorised into several different groups. Similarity measures can be broadly categorised into corpus-based and knowledge-based (Zhu and Iglesias, 2018). Corpus-based approaches use various collections of text as source whereas knowledge-based approaches incorporate ontologies like WordNet. How various resources are used for measuring the similarity is also used as the major component to categorise various measures (Altinel and Ganiz, 2018). Here, the similarity measures were grouped based on the knowledge bases, algorithms used in large text collections, machine learning, string matching techniques, and language syntaxes. Based on the ideas that have been used for measuring similarity between words using WordNet as knowledge source, the methods are categorised into one of the four measures viz path-based, information content-based, feature-based and hybrid measures (Qu et al., 2018).

Rest of the paper is organised as follows; Section 2 discusses about the background and importance of semantic similarity measure in various NLP applications that exploit corpus-based or knowledge-based measures. Section 3 introduces and explains the proposed approach in detail followed by experimental evaluations in Section 4. Section 5 concludes this work.

2 Background

Semantic similarity is one of the much needed measurements in various interrelated domains like NLP, artificial intelligence, information retrieval, sentiment analysis and more. They help us in quantifying the relationship between texts which in turn help us in applications including searching the documents, understanding the language, disambiguating words, identifying named entities and more. These measures make use of existing resources including domain dependent and domain independent text collections, dictionary sources, ontologies, web documents, etc., based on their applications. The research on semantic similarity is one of the well studied problem for English language. For other languages like, Arabic, Chinese, European languages and Indian languages the problem is yet to be studied well. The availability of language dependent well organised text collections, dictionary sources or ontologies for these languages is very limited. This is one of the reasons why the research on these languages is still open.

WordNet for English is a lexical database that arranged and stored English words under various language elements that helps the research in language processing. The contribution of WordNet in language research is a huge success. Hence, researchers have started building such ontology for multiple different languages in the world. Most of them have built language dependent ontology using links to English WordNet. For example, Arabic WordNet (Black et al., 2006), FinnWordNet (Lindén and Carlson, 2010), and WOLF (Sagot and Fišer, 2008) are some of the lexical databases built from Princeton WordNet. Indo WordNet (Bhattacharyya, 2010) is a lexical database of several Indian languages. Indo WordNet is constructed for different languages by building Hindi as a central linking language to Princeton WordNet.

Table 1 Literature on corpus-based, knowledge-based and hybrid semantic similarity measures

Method	*Language*	*Details of resources*		*Technique*
		Type	*Source*	
Lesk (1986)	English	Knowledge-based	WordNet	Any word overlapping
Adapted Lesk (Banerjee and Pedersen, 2002)	English	Knowledge-based	WordNet	Gloss overlapping
WSD in Arabic (Alkhatlan et al., 2018)	Arabic	Corpus-based	Arabic news, Arabic WordNet (AWN)	Word embedding
Hindi WSD (Singh et al., 2013)	Hindi	Knowledge-based	Hindi WordNet	Path length between noun concepts
WSD in cross-lingual IR (Thenmozhi and Aravindan, 2018)	Tamil – English	Knowledge-based	WordNet	Word overlapping
Textual similarity in Bengali text (Shahjalal and Aono, 2018)	Bengali	Corpus-based	Bengali Wikipedia	Word embedding
Three stage framework for Chinese word similarity (Huang et al., 2018)	Chinese	Corpus + knowledge-based	Web documents, Chinese WordNet	Word embedding, path length
Plagiarism detection in Malayalam language (Sindhu and Idicula, 2017)	Malayalam	Corpus-based	Web documents, vector space model	Jaccard, dice and cosine similarity
Context-based Arabic WSD (Bekkali and Lachkar, 2018)	Arabic	Corpus-based	Web documents, BableNet	Web-RST similarity measure (proposed)

With respect to the NLP and related domains, the existing similarity measuring approaches can be broadly categorised based on the type of data sources used for the calculation. They can be either corpus based, or knowledge based or hybrid of these two (Zhu and Iglesias, 2018). Table 1 shows various semantic similarity finding works that use either of corpus-based, knowledge-based or hybrid approaches on different languages.

2.1 Semantic similarity: corpus-based measures

The primary objective of corpus-based measures is to understand the associations between words and its contexts to measure the similarity. These measures exploit the information collected from large text collections to measure the similarity between two entities. The words extracted from the large text corpora and its frequency goes into a vector space to identify the hidden relationship between documents and the cosine angle between two vectors are used to measure the similarity in a corpus-based similarity approach called latent semantic analysis (LSA) (Landauer and Dumais, 1997). Word2Vec (Mikolov et al., 2013) is another similarity approach that constructs documents into vector space for similarity calculation. This approach implements the idea called word embedding. This idea became successful and being used for language processing widely. As it involves the statistics about the corpus or data collection to find the relationship within the text, it can be applied on any applications without the language barrier. Word embedding can be obtained using the definitions of each sense of words from Arabic WordNet and the Arabic words are disambiguated (Alkhatlan et al., 2018). Word embedding is used for finding similarity between Bengali text using Bengali Wikipedia as the data corpus (Shahjalal and Aono, 2018). A research on application of word embedding on Tamil language using content-based model proved performing well (Ajay et al., 2016).

2.2 Semantic similarity: knowledge-based measures

Knowledge-based measures use ontology to measure the similarity. If two words are placed near in ontology, then there is a possibility that these two words are similar. Usually, these types of similarity measures use the hierarchical knowledge and the textual content available through ontology to calculate the similarity values. WordNet (Miller, 1995) is one such ontology with rich contents to define a word in all possible ways. It links words hierarchically with other words under all appropriate part of speeches and various senses. That knowledge helps various researchers to use WordNet. As it is a source of hierarchical relations and textual information, various researches on similarity measurement exploits either of the contents. LCH (Leacock and Chodorow, 1998), WUP (Wu and Palmer, 1994), and HSO (Hirst and St-Onge, 1998) are few of the similarity measures that used the taxonomical knowledge of WordNet. Lesk (1986), and gloss vector (Patwardhan and Pedersen, 2006) uses the textual knowledge of WordNet in measuring the similarity. Extended Lesk (Banerjee and Pedersen, 2002) measure uses both the taxonomical and textual knowledge of WordNet. The taxonomical features of Hindi WordNet are used to find semantic relatedness between Hindi words in a proposal for Hindi sense disambiguation (Singh et al., 2013). Gloss overlapping between word definitions is also used for Hindi word sense disambiguation (Gautam and Sharma, 2016).

3 Architecture

Figure 1 shows the architecture of the proposed approach for calculating semantic similarity between Tamil words. This hybrid approach incorporates various dictionaries and the descriptions of words as extracted from these sources for finding the similarity.

3.1 Gloss definitions extraction

First, the gloss definitions and example sentences (if any) for the input words are extracted from the dictionary sources Indo Tamil WordNet (Kanojia et al., 2018), Oxford Tamil Dictionary, and Tamil Wikitionary. We have identified through our study that these are few of the reliable resources that defines Tamil words near to what WordNet does for English. Indo WordNet (Bhattacharyya, 2010) is a lexical knowledge source of Indian languages. It is like WordNet and EuroWordNet for English and European languages respectively. We used Indo WordNet for Tamil language. It consists of more than 25,000 synsets under various parts of speeches including noun, and verb along with gloss definitions and examples for each word. Like English WordNet, here too the words are arranged and linked with other related words through the relations like hypernyms, and hyponyms. This structure helps us in finding the relations of a word with other words and how close they are. The next resource, Oxford Tamil Dictionary is rich in defining Tamil words. It is being updated often after thorough study by experts in Tamil language. Wikitionary is another resource which is being used widely in NLP related tasks by researchers. It acts like a multilingual dictionary and provides Tamil meanings and definitions for various words under the project Tamil Wikitionary. We have used all the three resources to collect the gloss definitions and examples for the input words. The reason for using all these three resources is that each of these resources may define a word differently from the other and also we may get more example sentences that would give a concrete idea about understanding a word and its usage. The Oxford Tamil Dictionary and Indo WordNet Tamil both share mostly the common definitions.

We collect the gloss definitions and the example sentences from the above said resources. In reality, we have observed the following properties about the description of words by these resources:

1 a word may be defined using distinctive sentences that may have similar or different terms in each of these resources

2 not necessarily all of these resources have provided enough example sentences

3 each may have defined one or more senses of a given word

4 few of these resources may provide synonyms for the input words.

As we collect glosses from various resources, first it is mandatory to cleanse the definitions and identify the possible number of senses for each input word. To identify the possible senses of input word from various resources, we need to find the similarity between gloss definitions. This is achieved in this work using word overlapping. For that purpose, we have used Tamil Shallow Parser (http://ltrc.iiit.ac.in/analyzer/tamil/) for extracting the root words of noun and verb senses only from the gloss definitions in step 2 through step 4. As next problem, we found that some of the definitions are too short and may be ambiguous. Some of them have only synonyms and no definitions. For such words, we need to search further for getting some supportive definitions. We use the following steps to achieve the above said things for each input word:

Step 1 *Extract glosses from dictionary sources* – in this step, the direct glosses of input words and example sentences (if any) are extracted from dictionary sources directly. The extracted definitions and examples are stored separately for later use.
Table 2 explains an example gloss extraction for the word ‘இரத்தினம் (Irattiṉam)’. It shows the available senses for the input word, its gloss definitions for each such sense, and example sentences (if any) from different resources.
For our example input word ‘இரத்தினம் (Irattiṉam)’, all the three resources define only one noun sense. It is possible for each of these dictionary resources to define a word under various different senses. For example, the word ‘பழம் (Paḻam)’ is defined under only one noun sense in Indo Tamil WordNet, whereas it is defined under two noun senses in Oxford Tamil Dictionary and Tamil Wikitionary.

Step 2 *Identify senses for input word* – this step identifies different possible definitions of an input word from the dictionary sources and unifies them under appropriate senses based on the words present in the definitions. This is achieved by finding the word overlapping between definitions of each sense of different dictionary sources. If more than 50% of the words of a definition of one resource is used by the other resource, then both definitions are considered under same sense. For example, in Table 1, both Indo WordNet Tamil and Oxford Tamil Dictionary have shared most of the words in their definitions to define one sense. Hence, they both considered as the definitions of one particular sense. One interesting property between Indo WordNet Tamil and Oxford Dictionary is that they both share similar definitions than that of Tamil Wikitionary. In this step, the definitions with less than two keywords are not included for finding the word overlapping exercise. Hence, for time being the word with short definitions are considered under a separate sense. So far, we have identified and unified definitions under two senses as shown in the Table 3.

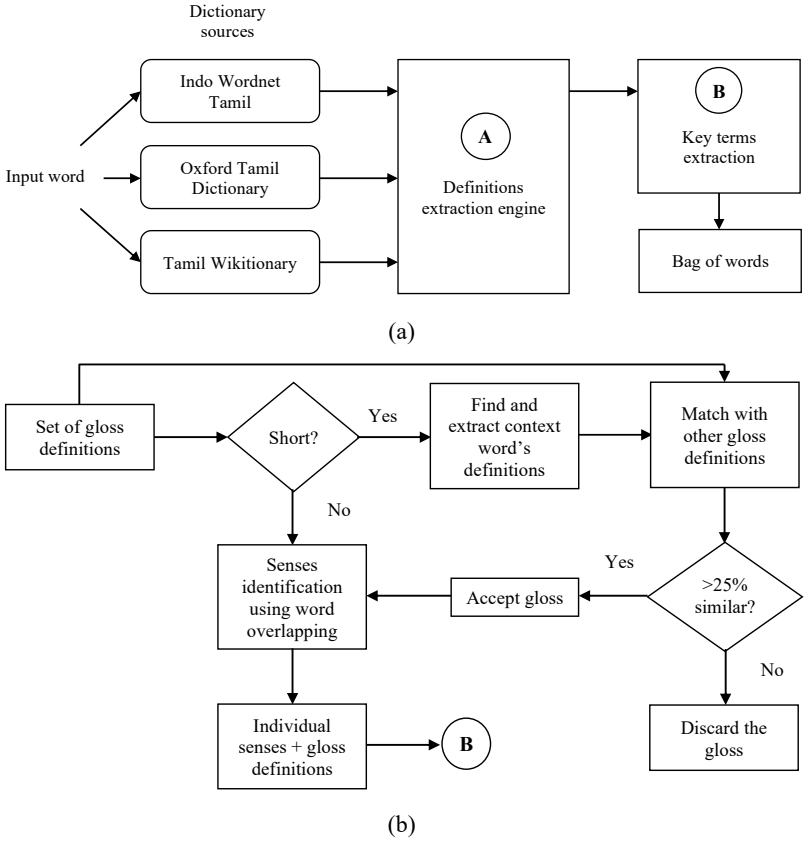
Step 3 *Identify the short gloss definitions* – the gloss definitions with less than two context keywords are considered as short definitions. This simple thumb rule is used to make a decision on short glosses. For example, in Table 2, the definition for ‘இரத்தினம் (Irattiṉam)’ by Tamil Wikitionary is a single word ‘மணி (Maṇi)’. This may not help in defining the input word much clear.

Step 4 *Identify, extract and filter the supportive definitions to choose right sense* – the short glosses would not help in targeting the right sense for the word defined. Hence, we search for the context word in the same resource further. For example, we search ‘மணி (Maṇi)’ in Tamil Wikitionary to get a concrete definition that would possibly support the previous definition. While we search for the word ‘மணி (Maṇi)’ in Tamil Wikitionary, we get the following definitions as given in Table 4.

Table 2 Gloss definitions and example sentences for various sense of input word ‘இரத்தினம் (Irattiṉam)’

Source	*Sense*	*Gloss definitions*	*Example sentences*
Indo WordNet Tamil	Noun 1	அணிகலன்களில் அழகுக்காகப் பதிக்கும் மரகதம், பவளம் போன்ற விலையுயர்ந்த இயற்கைப் பொருள் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta iyaṟkaip poruḷ)	சீமா இரத்தினமாலை அணிந்திருக்கிறாள் (Cīmā irattiṉamālai aṇintirukkiṟāḷ)
Oxford Tamil Dictionary	Noun 1	அணிகலன்களில் அழகுக்காகப் பதிக்கும் மரகதம், பவளம் போன்ற விலையுயர்ந்த ஒரு வகைக் கல் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta oru vakaik kal)	No example
Tamil Wikitionary	Noun 1	மணி (Maṇi)	No example

Figure 1 Proposed system architecture, (a) architecture layout (b) definitions extraction engine



Now, we got five senses of the context word 'மணி (Maṇi)'. All these senses may or may not be related to the input word 'இரத்தினம் (Irattiṉam)' either directly or indirectly. Hence, to find the association we find the word overlapping between each sense of context word with the finalised sense of input word. One important thing is that the context word helps in defining the input word but not equal to the input word. Hence, we check for a match of keywords of context word definitions with each of the finalised sense of input word definitions. If any one word matches with the definitions of finalised word, then we conclude that that particular sense of the context word is used as the supporting definition for the short definitions. In our example, it is found that the definition of sense noun 1 shares most words with that of noun 1 sense of input word 'இரத்தினம் (Irattiṉam)'. So, we conclude that noun 1 of 'மணி (Maṇi)' is the most relevant definition. This definition is concatenated with the short definition to produce a new definition. For our example, 'மணி (Maṇi)' is expanded as 'மணி ஒளி வீசும் வைரம் போன்ற கல் (Maṇi Oḷi vīcum vairam pōṉṟa kal)'. This was found as the most supporting definition using the overlapping with the finalised noun 1 sense of input word 'இரத்தினம் (Irattiṉam)', the new definition also unified under sense noun 1 of input word. That is, we conclude that the word 'இரத்தினம் (Irattiṉam)' has only one noun sense with three definitions as shown in Table 5.

Table 3 Identified and unified senses of input word using different dictionary resources for 'இரத்தினம் (Irattiṉam)'

Sense	*Gloss definitions*	*Example sentences*
Noun 1	• அணிகலன்களில் அழகுக்காகப் பதிக்கும் மரகதம், பவளம் போன்ற விலையுயர்ந்த இயற்கைப் பொருள் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta iyaṟkaip poruḷ) • அணிகலன்களில் அழகுக்காகப் பதிக்கும்) மரகதம், பவளம் போன்ற விலையுயர்ந்த ஒரு வகைக் கல் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta oru vakaik kal)	சீமா இரத்தினமாலை அணிந்திருக்கிறாள் (Cīmā irattiṉamālai aṇintirukkiṟāḷ)
Noun 2	• மணி (Maṇi)	No example

Table 4 Example definitions for the context word 'மணி (Maṇi)' using Tamil Wikitionary

Sense	*Gloss definitions*	*Example sentences*
Noun 1	ஒளி வீசும் வைரம் போன்ற கல் (Oḷi vīcum vairam pōṉṟa kal)	No example
Noun 2	அழகு (Aḻaku)	மணி மணியாக எழுதுதல் (Maṇi maṇiyāka eḻututal)
Noun 3	பாசி (Pāci)	No example
Noun 4	வெண்கலத்தால் செய்த, நடுவில் நாவுடன் அமைந்து, ஆட்டினால் ஒலி எழுப்பும் கவிழ்ந்த கிண்ணம் போன்ற கருவி (Veṇkalattāl ceyta, naṭuvil nāvuṭaṉ amaintu, āṭṭiṉāl oli eḻuppum kaviḻnta kiṇṇam pōṉṟa karuvi)	No example
Noun 5	60 நிமிட கால அளவு.காலத்தைக்குறிக்கும் பெயர் (60 Nimiṭa kāla aḷavu.Kālattaikkuṟikkum peyar)	No example

Table 5 Final set of sense and definitions for further action for the input word 'இரத்தினம் (Irattiṉam)'

Sense	*Gloss definitions*
Noun 1	• அணிகலன்களில் அழகுக்காகப் பதிக்கும் மரகதம், பவளம் போன்ற விலையுயர்ந்த இயற்கைப் பொருள் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta iyaṟkaip poruḷ) • அணிகலன்களில் அழகுக்காகப் பதிக்கும்) மரகதம், பவளம் போன்ற விலையுயர்ந்த ஒரு வகைக் கல் (Aṇikalaṉkaḷil aḻakukkākap patikkum marakatam, pavaḷam pōṉṟa vilaiyuyarnta oru vakaik kal) • மணி ஒளி வீசும் வைரம் போன்ற கல் (Maṇi Oḷi vīcum vairam pōṉṟa kal)

To find the most supporting definition of the context word through word overlapping, we have also included the example sentences of input word to find any match. Further, if no match is found after the application of the example sentences, we simply discard the short definition in question. Now what we have is the gloss definitions and example sentences for each identified sense of input word.

The algorithm *AlternateGlossForShortGloss* is used for identifying and extracting alternate gloss definitions that support short glosses.

Algorithm *AlternateGlossForShortGloss*

Input: Definitions of short gloss keywords (SG), gloss definitions from dictionary sources (G)

Output: Alternate gloss definitions (A)

```
1:  K_sg, K_g, A ← {∅}
2:  for each gloss sg ∈ SG do
3:      K_sg ← K_sg ∪ Extract_Keywords(sg)
4:  end for
5:  for each gloss g ∈ G do
6:      K_g ← K_g ∪ Extract_Keywords(g)
7:  end for
8:  for each gloss sg ∈ SG do
9:      for each gloss g ∈ G do
10:         similarity ← Intersect(K_sg, Kg)
            / Union(K_sg, K_g)
11:         if similarity > 0.25 then
12:             A ← A ∪ gloss sg
13:         else
14:             discard gloss
15:         end if
16:     end for
17: end for
```

The algorithm *SensesDeclaration* is for finding the relevancy of senses from one source dictionary source to the other to group gloss definitions into single sense.

Algorithm *SensesDeclaration*

Input: Keyword sets of each sense of input word from sources A and B

Output: Gloss definitions under various senses

```
1:  MAX ← Senses_Count(A) + Senses_Count(B)
2:  new_senses[MAX] ← {∅}
3:  sense_number ← 0
4:  for each sense keyword set a ∈ A do
5:      sense_number++
6:      new_senses[sense_number] ← a
7:  end for
8:  for each sense keyword set b ∈ B do
9:      sense_number++
10:     new_senses[sense_number] ← b
11: end for
12: sense_number ← 0
13: for each sense keyword set a ∈ A do
14:     for each sense keyword set b ∈ B do
15:         similarity ← Intersect(a, b) / Union(a, b)
16:         if similarity > 0.5 then
17:             sense_number++
18:             new_senses[sense_number] ←
                new_senses[sense_number] ∪ b
19:             new_senses[Senses_Count (A)
                + b] ← {∅}
20:         end if
21:     end for
22: end for
```

Table 6 Similarity scores calculated using different approaches on Miller Charles noun pair dataset

Miller and Charles noun pairs		*Noun pairs in Tamil*		*Our method*	*Adapted Lesk*	*Miller Charles*	*RG*	*Gloss vector*
Gem	Jewel	இரத்தினம் (Rattiṉam)	அணிகலன்/நகை (Aṇikalaṉ/Nakai)	0.25	816	3.84	3.91	1.000
Journey	Voyage	பயணம் (Payaṇam)	கடற்பயணம் (Kaṭaṟpayaṇam)	0.25	41	3.84	3.58	0.204
Boy	Lad	பையன்/சிறுவன் (Paiyaṉ/Ciṟuvaṉ)	இளைஞன் (Iḷaiñaṉ)	0.33	50	3.76	3.82	0.736
Coast	Shore	கடற்கரை (Kaṭaṟkarai)	கரை (Karai)	0.02	51	3.70	3.60	0.643
Magician	Wizard	மந்திரவாதி (Mantiravāti)	மாந்திரிகன் (Māntirikaṉ)	0.70	145	3.50	3.21	1.000
Midday	Noon	நண்பகல் (Naṇpakal)	மத்தியானம் (Mattiyāṉam)	1.00	46	3.42	3.91	1.000
Furnace	Stove	உலை/உலைஅடுப்பு (Ulai/Ulai'aṭuppu)	அடுப்பு (Aṭuppu)	0.65	52	3.11	3.11	0.576
Food	Fruit	உணவு (Uṇavu)	பழம் (Paḻam)	0.33	34	3.08	2.69	0.298
Bird	Cock	பறவை (Paṟavai)	சேவல் (Cēval)	0.63	115	3.05	2.63	0.659
Bird	Crane	பறவை (Paṟavai)	நாரை (Nārai)	0.67	18	2.97	2.63	0.358
Brother	Monk	சகோதரன் (Cakōtaraṉ)	துறவி (Tuṟavi)	0.23	147	2.82	2.74	0.429
Lad	Brother	இளைஞன் (Iḷaiñaṉ)	சகோதரன் (Cakōtaraṉ)	0.07	10	1.66	2.41	0.424
Journey	Car	பயணம் (Payaṇam)	ஊர்தி (Ūrti)	0.21	19	1.16	1.55	0.327
Monk	Oracle	துறவி/சன்னியாசி (Tuṟavi/Caṉṉiyāci)	வாக்கு/அசரீரி (Vākku/Acarīri)	0.22	4	1.1	0.91	0.129
Cemetery	Woodland	மயானம் (Mayāṉam)	வனம்/காடு (Vaṉam/Kāṭu)	0.05	7	0.95	1.18	0.077
Food	Rooster	உணவு (Uṇavu)	சேவல் (Cēval)	0.05	8	0.89	1.09	0.115
Coast	Hill	கடற்கரை/கரை (Kaṭaṟkarai/Karai)	குன்று/மலை (Kuṉṟu/Malai)	0.1	11	0.87	1.26	0.233
Forest	Graveyard	காடு (Kāṭu)	இடுகாடு (Iṭukāṭu)	0.04	7	0.84	1.00	0.089
Shore	woodland	கரை (Karai)	வனம்/காடு (Vaṉam/Kāṭu)	0.14	4	0.63	0.90	0.103
Monk	Slave	துறவி (Tuṟavi)	அடிமை (Aṭimai)	0.12	13	0.55	0.57	0.249
Coast	Forest	கடற்கரை/கரை (Kaṭaṟkarai/Karai)	காடு (Kāṭu)	0.11	10	0.42	0.85	0.163
Lad	Wizard	இளைஞன் (Iḷaiñaṉ)	மாந்திரிகன் (Māntirikaṉ)	0.03	2	0.42	0.99	0.042
Chord	Smile	நாண் (Nāṇ)	புன்னகை (Puṉṉakai)	0	1	0.13	0.02	0.063
Glass	Magician	கண்ணாடி (Kaṇṇāṭi)	மந்திரவாதி (Mantiravāti)	0.01	3	0.11	0.44	0.068
Rooster	Voyage	சேவல் (Cēval)	கடற்பயணம் (Kaṭaṟpayaṇam)	0	1	0.08	0.04	0.026
Noon	String	நண்பகல் (Naṇpakal)	சரம் (Caram)	0	2	0.08	0.04	0.000
Car	Automobile	சிற்றுந்து/கார் (Ciṟṟuntu/Kār)	மோட்டார்-வாகனம் (Mōṭṭārvākaṉam)	1	3576	3.92	3.92	1
Crane	Implement	பாரம்தூக்கும்இயந்திரம் (Pāramtūkkumiyantiram)	கருவி (Karuvi)	0.78	6	1.68	2.37	0.162

3.2 Extract key terms

The sentences including definitions and examples that are extracted from different resources are now parsed to identify and extract key terms. The extracted sentences were parsed using a shallow parser for Tamil which was developed by the Language Technologies Research Center, IIIT Hyderabad. This parser works at various levels including morphological analysis, tagging and chunking to produce the final parsed text content. This proposal is towards finding similarity between noun words and hence

we are interested and have extracted only the noun and verb key terms that are used in defining the given word from the definitions. The extracted key terms are collected in a bag that represents one input word. Let us take the gloss definition 'பொதுவாக ஆண் பறவை; குறிப்பாகக் கோழிகளில் ஆண் இனம்' (Potuvāka āṇ paṟavai; kuṟippākak kōḻikaḷil āṇ iṉam) for the input word 'சேவல் (Cēval)'. We get the following keywords post parsing of this statement; "ஆண் (Āṇ)", "பறவை (Paṟavai)", "கோழி (Kōḻi)", and "இனம் (Iṉam)". It can be observed from the above list that the collected keywords are the root words of the keywords present in the gloss definition. The keywords present in the sentence may not be used as they are while finding similarity due to the inflected nature of Tamil words. The keywords that are collected in this way are put in separate bags for every sense of each input word.

3.3 Similarity calculation

The similarity between both the input Tamil words is found using the Jaccard similarity coefficient. The word overlapping between each individual sense of both input words are found and the Jaccard similarity is calculated. The Jaccard measure is used to find how similar two sets (bag of words) are. The measure is defined as the total number of common elements between two sets divided by the total number of elements present in two sets.

$$Jaccard(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

Here, W_1 and W_2 represent the bag of words for both input words respectively. The highest similarity value between any two senses of two different input words are considered as the similarity score between those words.

The proposed calculation involves the following weighting methods to boost up and normalise the similarity score.

- *Scheme 1:* if one input word is present as the synonym of the other input word, the weight *w* will be 1
- *Scheme 2:* if one input word is the hypernym, or hyponym of the other input word (applicable only for the resource Indo WordNet), the weight *w* will be 1
- *Scheme 3:* if one input word is present in the gloss of the other input word, the weight *w* will be 0.5.

According to the schemes listed above, the final similarity value will be calculated through normalisation as follows:

$$Sim(W_1, W_2) = \frac{Jaccard(W_1, W_2) + w}{Number\_of\_presence + 1}$$

In the above equation, the *Number_of_presence* is the presence of input word according to various schemes listed above. For example, if the input word is present according to scheme 1 and scheme 2, then the *number_of_presence* will be 3. The calculated similarity score for Miller Charles dataset are shown in Table 6.

4 Experimental evaluations

4.1 Datasets

For Tamil, the similarities between words have not been tried before. Hence, we do not have a dataset for finding semantic relatedness in Tamil. Miller and Charles (1991) and Rubenstein and Goodenough (1965) are the two of the gold standard datasets for finding semantic relatedness between English words. These two datasets share the similarity values between words that are evaluated by human evaluators. We have translated the words from these datasets to our requirement. The direct translation were not working for all the words due to one or more of the following reasons:

a direct translation of the input word from these datasets to Tamil is not available

b some translation resulted in few words or sentence

c in few cases both the input words translated to same word.

Hence, we have used native Tamil speakers to conclude the correct translation of these words. The native speakers are suggested for this kind of translation in literature (Akhtar et al., 2017; Camacho-Collados et al., 2015; Dai and Huang, 2011; Freitas et al., 2016). This resulted in removal of few words. The datasets are described in Table 6.

The results produced by the proposed approach is compared with the human evaluated datasets including Miller and Charles and Rubenstein Goodenough and with the other well known gloss-based similarity measures including adapted Lesk and gloss vector are furnished in Table 7.

The graphs shown in Figure 2 give the comparison of our method with other known methods.

Table 7 Pearson correlation coefficient for various similarity approaches

Methods	*Type*	*Miller Charles*	*Rubenstein Goodenough*
Proposed	Overlapping of dictionary glosses	0.6568	0.7783
Gloss vector	Cosine angle between WordNet glosses	0.8712	0.8611
Adapted Lesk	Gloss overlapping	0.8432	0.8421

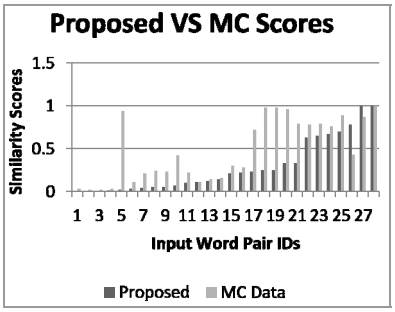
4.2 Evaluation

We do not have any dataset to test against Tamil words similarity. Hence, we used the results produced for English word datasets. To evaluate the performance of the proposed method, we have used the Pearson correlation coefficient to study the relationship among the proposed result against the human evaluated scores as per Miller Charles and

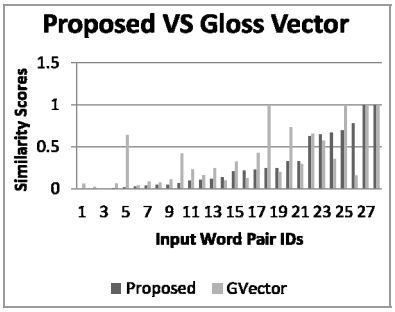
Rubenstein Goodenough. The Pearson coefficient (r) analyses the linear relationship between two input sets and is calculated as follows:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

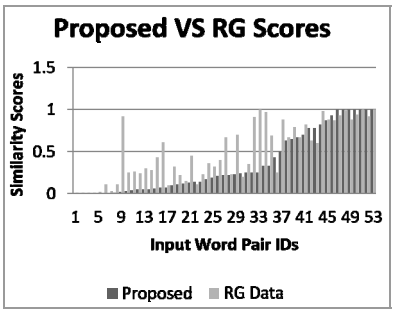
Figure 2 Comparison of similarity scores of the proposed method against, (a) human judged scores of Miller Charles dataset (b) gloss vector scores of Miller Charles dataset (c) human judged scores of Rubenstein Goodenough dataset (d) gloss vector scores of Rubenstein Goodenough dataset



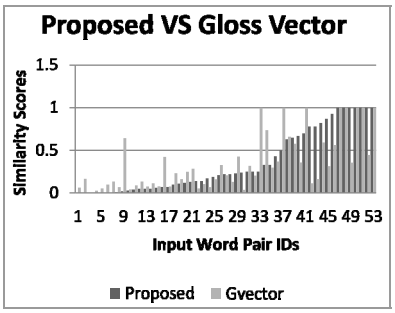
(a)



(b)



(c)



(d)

The Pearson's coefficient of various approaches for Miller Charles and Rubenstein Goodenough datasets along with the proposed one are listed in Table 7.

5 Conclusions

The similarity between words would help in various NLP and artificial intelligence applications. In this work, an approach is proposed to measure the semantic similarity between Tamil words. This work has been taken due to the unavailability of such measures in Tamil language. The proposed approach is an unsupervised approach that exploits the existing dictionary sources for Tamil language as major data source. Major contribution for the approach is taken from Indo WordNet for Tamil, an ontology-based thesaurus for Tamil language. The approach exploits the word descriptions and example sentences from these resources in similarity calculation. The evaluation of the performance of our approach on Miller and Charles and Rubenstein and Goodenough datasets using Pearson correlation proved the ability of our approach in par with other similar approaches in English language. Though the proposed approach performed well in recognising the similarity between two words, the lack of definitions for few word senses degrades the overall performance. If we would construct a complete ontology for Tamil like WordNet for English, the proposed method can perform well.

References

Ajay, S.G., Srikanth, M., Kumar, M.A. and Soman, K.P. (2016) 'Word embedding models for finding semantic relationship between words in Tamil language', *Indian Journal of Science and Technology*, Vol. 9, No. 45, pp.1–5.

Akhtar, S.S., Gupta, A., Vajpayee, A., Srivastava, A. and Shrivastava, M. (2017) 'Word similarity datasets for Indian languages: annotation and baseline systems', in *Proceedings of the 11th Linguistic Annotation Workshop*, April, pp.91–94.

Akmal, S., Shih, L.H. and Batres, R. (2014) 'Ontology-based similarity for product information retrieval', *Computers in Industry*, Vol. 65, No. 1, pp.91–107.

Alenazi, M., Reddy, D. and Niu, N. (2018) 'Assuring virtual PLC in the context of SysML models', in *International Conference on Software Reuse*, Springer, Cham, May, pp.121–136.

Alkhatlan, A., Kalita, J. and Alhaddad, A. (2018) 'Word sense disambiguation for arabic exploiting Arabic WordNet and word embedding', *Procedia Computer Science*, Vol. 142, pp.50–60.

Altınel, B. and Ganiz, M.C. (2018) 'Semantic text classification: a survey of past and recent advances', *Information Processing & Management*, Vol. 54, No. 6, pp.1129–1153.

Azad, H.K. and Deepak, A. (2019) *A New Approach for Query Expansion Using Wikipedia and WordNet*, arXiv preprint arXiv: 1901.10197.

Banerjee, S. and Pedersen, T. (2002) 'An adapted Lesk algorithm for word sense disambiguation using WordNet', in *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, February, pp.136–145.

Bekkali, M. and Lachkar, A. (2018) 'Context-based Arabic word sense disambiguation using short text similarity measure', in *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications*, ACM, October, pp.1–6.

Bhattacharyya, P. (2010) 'Indowordnet', *Lexical Resources Engineering Conference 2010 (LREC 2010)*, Malta, May, pp.3785–3792.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006) 'The Arabic WordNet project', in *Proceedings of LREC 2006*, pp.29–34.

Camacho-Collados, J., Pilehvar, M.T. and Navigli, R. (2015) 'A framework for the construction of monolingual and cross-lingual word similarity datasets', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2: Short Papers, pp.1–7.

Dai, L. and Huang, H. (2011) 'An English-Chinese cross-lingual word semantic similarity measure exploring attributes and relations', in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pp.467–476.

Ehsan, N. and Shakery, A. (2016) 'Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information', *Information Processing & Management*, Vol. 52, No. 6, pp.1004–1017.

Freitas, A., Barzegar, S., Sales, J.E., Handschuh, S. and Davis, B. (2016) 'Semantic relatedness for all (languages): a comparative analysis of multilingual semantic relatedness using machine translation', in *European Knowledge Acquisition Workshop*, Springer, Cham, November, pp.212–222.

Gabsi, I., Kammoun, H. and Amous, I. (2017) 'MeSH-based disambiguation method using an intrinsic information content measure of semantic similarity', *Procedia Computer Science*, Vol. 112, pp.564–573.

Gautam, C.B.S. and Sharma, D.K. (2016) 'Hindi word sense disambiguation using Lesk approach on bigram and trigram words', in *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, ACM, August, p.81.

Hirst, G. and St-Onge, D. (1998) 'Lexical chains as representations of context for the detection and correction of malapropisms', *WordNet: An Electronic Lexical Database*, Vol. 305, pp.305–332.

Huang, D., Pei, J., Zhang, C., Huang, K. and Ma, J. (2018) 'Incorporating prior knowledge into word embedding for Chinese word similarity measurement', *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 17, No. 3, pp.1–21.

Kanojia, D., Patel, K. and Bhattacharyya, P. (2018) 'Indian language WordNets and their linkages with Princeton WordNet', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pp.4603–4606.

Landauer, T.K. and Dumais, S.T. (1997) 'A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge', *Psychological Review*, Vol. 104, No. 2, pp.211–240.

Leacock, C. and Chodorow, M. (1998) 'Combining local context and WordNet similarity for word sense identification', *WordNet: An Electronic Lexical Database*, Vol. 49, No. 2, pp.265–283.

Lesk, M. (1986) '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*, ACM, June, pp.24–26.

Lindén, K. and Carlson, L. (2010) *FinnWordNet-WordNet på finska via översättning*, Vol. 17, pp.119–140, LexicoNordica.

Mezzoudj, F. and Benyettou, A. (2018) 'An empirical study of statistical language models: n-gram language models vs. neural network language models', *International Journal of Innovative Computing and Applications*, Vol. 9, No. 4, pp.189–202.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality', in *Advances in Neural Information Processing Systems*, pp.3111–3119.

Miller, G.A. (1995) 'WordNet: a lexical database for English', *Communications of the ACM*, Vol. 38, No. 11, pp.39–41.

Miller, G.A. and Charles, W.G. (1991) 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, Vol. 6, No. 1, pp.1–28.

Nasir, J.A., Varlamis, I. and Ishfaq, S. (2019) 'A knowledge-based semantic framework for query expansion', *Information Processing & Management*, Vol. 56, No. 5, pp.1605–1617.

Patwardhan, S. and Pedersen, T. (2006) 'Using WordNet-based context vectors to estimate the semantic relatedness of concepts', in *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*.

Pera, M.S. and Ng, Y.K. (2007) 'Using word similarity to eradicate junk emails', in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, ACM, November, pp.943–946.

Qu, R., Fang, Y., Bai, W. and Jiang, Y. (2018) 'Computing semantic similarity based on novel models of semantic representation using Wikipedia', *Information Processing & Management*, Vol. 54, No. 6, pp.1002–1021.

Rubenstein, H. and Goodenough, J.B. (1965) 'Contextual correlates of synonymy', *Communications of the ACM*, Vol. 8, No. 10, pp.627–633.

Sagot, B. and Fišer, D. (2008) 'Building a free French WordNet from multilingual resources', in *OntoLex*, May, pp.1–6.

Shahjalal, M. and Aono, M. (2018) 'Semantic textual similarity in Bengali text', *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, pp.1–5.

Sindhu, L. and Idicula, S.M. (2017) 'Plagiarism detection in Malayalam language text using a composition of similarity measures', in *Proceedings of the 9th International Conference on Machine Learning and Computing*, ACM, February, pp.456–460.

Singh, S., Singh, V.K. and Siddiqui, T.J. (2013) 'Hindi word sense disambiguation using semantic relatedness measure', in *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, Springer, Berlin, Heidelberg, December, pp.247–256.

Thenmozhi, D. and Aravindan, C. (2018) 'Ontology-based Tamil-English cross-lingual information retrieval system', *Sādhanā*, Vol. 43, No. 10, pp.1–14.

Uddin, M.N., Duong, T.H., Nguyen, N.T., Qi, X.M. and Jo, G.S. (2013) 'Semantic similarity measures for enhancing information retrieval in folksonomies', *Expert Systems with Applications*, Vol. 40, No. 5, pp.1645–1653.

Wu, Z. and Palmer, M. (1994) 'Verbs semantics and lexical selection', in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, June, pp.133–138.

Zhu, G. and Iglesias, C.A. (2018) 'Exploiting semantic similarity for named entity disambiguation in knowledge graphs', *Expert Systems with Applications*, Vol. 101, pp.8–24.