INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Optimized short text embedding for bilingual similarity using Probase and BabelNet

Natasha J.
natashaj600@gmail.com
Anna University, CEG Campus, Chennai, Tamil Nadu

Vijayarani J.
viji.cs66@gmail.com
Anna University, CEG Campus, Chennai, Tamil Nadu

ABSTRACT

Most existing methodologies for text classification represent text as vectors of words, to be specific "bag-of-words." This content portrayal results in a high dimensionality of feature space and much of the time experiences surface jumbling. When it comes to short texts, these become even more serious because of their shortness and sparsity and with the bilingual similarity of text it gets more difficult. This paper proposes an approach to deal with both sparsity and computational complexity of bilingual similarity of short text. English short text is mapped with Probase and Hindi short text is mapped with BabelNet a knowledge base with coverage of words and concepts for 248 languages. A semantic network is created to manipulate the word to word and concept to concept correlation. Unlike the earlier approaches of embedding, words and concepts from both English and Hindi short texts are treated separately to yield word embedding (Word2Vec) and concept embedding (Concept2Vec) respectively. The similarity between bilingual short texts is computed using the skip-gram based word embedding and concept embedding. When evaluated with Pilot and STSS 131 short text benchmark datasets, the proposed optimized bilingual short text embedding gives better similarity score.

Keywords— *Short text, Conceptualization, Probase, BabelNet, Skipgram, Word2Vec, Concept2Vec*

1. INTRODUCTION

Many Natural Language Processing (NLP) applications require the input text to be represented as a fixed-length feature, of which the short-text embedding is very important, perhaps the most common fixed length vector representation for the text is the bag-of-words or the bag-ofn-grams, However they suffer severely from the data sparsity and high dimensionality and have very little sense about the semantics of the words. Recently, for short text representation and classification the Deep Neural Network (DNN) approaches have achieved the state-of-the-art results.

Despite their usefulness, the recent embedding of short-text faces several challenges: the first most short-text embedding models represent each short text only by using the words ' literal meanings, which makes these models indiscriminate to the ubiquitous polysemy. Second, for short text, neither parsing nor modeling of the topic works well because the input simply does not contain enough signals. In order to resolve this, more semantic signals (e.g. the concepts) must be derived from the short text input and enable the full interplay of the signals to generate a solid semantic representation for the given short text. Vector Space Models (VSM) are being used recently to represent text as a vector. The issue is VSMs are sparse and require dimensionality reduction. Recently, word embedding vector representation is more popular. Some of the variations of word embeddings (Word2Vec) are Sent2Vec, Phrase2Vec, Text2Vec, Doc2Vec etc. Average of embeddings of words in the short text will give the short text embedding. Similarity computed with only words in the short text does not provide complete semantic information. The solution is conceptualization with a knowledge base like Probase. Now it is possible to get the conceptualized short text embedding. Most of the earlier approaches have treated words and concepts in a similar manner. However, during the learning process of word embedding model words and concepts must be handled differently. Concept2Vec is useful for earning concept embedding which is different from word embedding Concept2Vec Skip gram model is capable of predicting the contextually related concepts for the given concept which will be useful in discriminating the concept embedding from word embedding. BabelNet is a multilingual lexicalized semantic network and ontology which has a total of 284 languages in the BabelNet version 4.0. Therefore, this study investigates how to obtain similarity for bilingual short text, words. A different knowledge base for English short text called Probase and different one called BabelNet for Hindi short text. BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms and a semantic network which connects concepts and named entities in a very large network of semantic relations, made up of about 16 million entries, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages.

Probase enables machines to better understand human communication. For example, in natural language processing

and speech analysis, knowledge bases can help reduce the ambiguities in language. As Probase has a knowledgebase as large as the concept space (of worldly facts) in a human mind, it has unique advantages in these applications. Besides, with the probabilistic knowledge provided by Probase, typicality score of the concepts can be computed.

In this project "Optimized Short Text Embedding For Bilingual Similarity Using Probase And Babelnet" a different approach to embedding is proposed where both words and concepts from the bilingual short text are treated separately by making use of word and concept vectorization and finally calculating the similarity of short texts from the average of vectors obtained from word and concept embedding.

2. RELATED WORK

2.1 Bilingual short text similarity

Ta-Chung Chi et al [4] introduced the first dataset for evaluating English-Chinese Bilingual Contextual Word Similarity, namely BCWS1. The dataset consists of 2,091 English-Chinese word pairs with the corresponding sentential contexts and their similarity scores annotated by the human. The annotated dataset has higher consistency compared to other similar datasets. They established several baselines for the bilingual embedding task to benchmark the experiments. Modeling cross-lingual sense representations as provided in this dataset has the potential of moving artificial intelligence from monolingual understanding towards multilingual understanding. To establish the Bilingual Contextual Word Similarity (BCWS) dataset, they collected the data through implementing a five steps procedure:

(a) Chinese Multi-Sense Word Extraction
(b) English Candidate Word Extraction
(c) Enriching Semantic Relationship
(d) Adding Contextual Information
(e) Human Labeling.

Their collected BCWS dataset includes 2,091 questions, each of which contains exactly one Chinese sentence and one English sentence Goran Glavaša et al [5] proposed an unsupervised and highly resource-light approach to measuring semantic similarity in different languages between texts. They projected continuous word vectors (i.e. word embedding) from one language to the vector space of the other language via the linear translation model to operate in the bilingual (or multilingual) space. They then align words in the bilingual embedding space according to the similarity of their vectors and investigate various unsupervised semantin similarity measures that exploit bilingual embedding and word alignment. The proposed approach is applicable to virtually any pair of languages for which only a limited set of word translation pairs between the languages are showing stability across different language pairs. In addition, they evaluated their method for two extrinsic tasks, namely the extraction of parallel sentences from comparable corporate and cross-lingual plagiarism detection they demonstrated that it delivers comparable performance to those of complex state of the art resource intensive models for the respective tasks.

Kerstin Denecke [6] introduced a methodology within a multilingual framework to determine text polarity. The method leverages the lexical resources available in English (SentiWordNet) for sentiment analysis. The method was tested for Amazon selected German film reviews and was compared to ann-grams-based statistical polarity classifier. The results showed that working in a multilingual framework with standard technology using existing approaches to sentiment analysis is a viable approach. The methods proposed to classify news documents in English with an accuracy between 51% and 62% depending on their polarity. German film reviews were slightly better classified with accuracies ranging from 58% to 66%. The results suggested that the accuracy of the various methods is not dependent on the language and domain being processed.

Edi Faisal et al [7] proposed the use of SVM algorithm with TF-IDF for feature extraction and Wikipedia as the training data to solve the Indonesian language WSD problem. Their proposed method results reached a level of accuracy of 0.877. This accuracy was considered good since there was no other method proposed for word sense disambiguation in Bahasa Indonesia. They divided their method process into five parts; (i) Wikipedia articles selection; (ii) training and testing data; (iii)Preprocessing; (iv)Machine Learning using SVM; (v) evaluating with a performance measure.

2.2 Probase: a probabilistic taxonomy for text understanding

W. Wu, et al (2012) [1], proposed the Probase taxonomy which is unique in three aspects: was used to extract isA pairs from web text and a taxonomy construction algorithm was used to connect these pairs into a hierarchical structure. The resulting taxonomy has the highest precision (92.8%). It is the first general-purpose taxonomy that takes a probabilistic approach to model the knowledge it possesses. Each fact or relation is associated with some probabilities to measure its plausibility and typicality. Plausibility is useful for detecting errors and integrating heterogeneous knowledge sources, while typicality is useful for conceptualization and inference. Probabilistic treatment allows Probase to better capture the semantics of human languages. It also has tens of thousands of specific concepts such as "renewable energy technologies", "meteorological phenomena" and "common sleep disorders", which cannot be found in Freebase or any other taxonomies.

2.2.1 Taxonomy construction: After extraction of information and cleaning, a large set of isa pairs is produced. Each pair represents an edge in the taxonomy. They constructed a taxonomy from these individual edges. The taxonomy was modeled as a DAG (Directed Acyclic Graph). A node in the taxonomy is either a concept node (e.g., company) or an instance node (e.g., Microsoft). A concept contains a set of instances and possibly a set of sub-concepts. An edge (u,v) connecting two nodes u and v mean that u is a super-concept of v. Differentiating concept nodes from instance nodes is natural in taxonomy.
The taxonomy inference framework was implemented on a cluster of servers using the Map-Reduce model. Probase was used in a graph database system called Trinity. They extracted 326,110,911 sentences from a corpus containing 1,679,189,480 web pages. The inferred taxonomy has 2,653,872 distinct concepts, 16,218,369 distinct concept-instance pairs and 4,539,176 unique concept-subconcept pairs (20,757,545 pairs in total). Next, the characteristics of the concept space and the ISA relationship space of Probate were analyzed and briefly evaluated for several applications that leverage typicality in conceptualization.

Computing semantic similarities between two terms are essential for a variety of applications for text analysis and understanding. Existing approaches, however, are more appropriate for word-semantic similarity rather than more general Multi-Word Expressions (MWEs) and they are not very well-scale. Peipei Li et al [3] proposed a lightweight and effective approach to semantic similarity using a large-scale semantic network acquired automatically from trillions of web documents. They then mapped them into the concept space with two terms and

compared their similarity. In addition, they also introduced a clustering approach to orthogonalize the concept of space to improve similarity accuracy. To compute the similarity between two terms, they computed the similarity between their contexts. The context they used comes from a large-scale, probabilistic semantic network, known as Probase. Given a term, they defined its concept context as the whole set of concepts in Probate to which the term belongs. Also, they performed concept clustering for automatic sense disambiguation. As an optimization, they pruned all the irrelevant clusters. Lastly, they defined the similarity of two terms as the highest similarity between any first term meaning and any second term meaning.

2.3 Babelnet

Roberto Navigli et al [8] presented a very large, broad-based, multilingual semantic network. The resource is built automatically through a methodology that integrates WordNet and Wikipedia lexicographic and encyclopedic knowledge. Furthermore, Machine Translation is also used to enrich the resource for all languages with lexical information. They conducted experiments on new and existing gold standard datasets to demonstrate the high quality and resource coverage. Key to their approach was to draw up a mapping between a multilingual repository of encyclopedic knowledge (Wikipedia) and an English computational lexicon (WordNet). There are several advantages to this integration process. First, the two resources contribute to different types of lexical knowledge, one is mostly concerned with named entities, and the other is concerned with concepts. Second, while Wikipedia is less structured than WordNet, it offers 223 large amounts of semantic relationships. Furthermore, they also contributed a wide range of sensory events harvested from Wikipedia and SemCor, a corpus that wa3ews entered into a state - of - the - art machine translation system to fill the gap between resource-rich languages, such as English, and resource-poor ones.

2.4 Concept2vec

Ontological concepts play a crucial role in the graphs of knowledge, providing them with high-quality embedding. Faisal Alshargi et al [9] introduced a framework containing three distinct tasks related to the individual aspects of ontological concepts, (i) the aspect of categorization, (ii) the hierarchical aspect, and (iii) the relational aspect. Then, a number of intrinsic metrics were proposed for each task to evaluate the quality of embedding. They also prepared an appropriate data set and conducted a series of comparative studies on popular embedding models for ontological concepts. In addition to a comparison of the quality of the available embedding models, w.r.t. multiple experimental studies were conducted on the framework. Using this framework they improved comparisons of ontological concepts with embedding. Ontological concepts play a crucial role in (i) capturing the semantics of a particular domain, (ii) typing entities that bridge a schema level and instance level, and (iii) identifying valid types of sources and destinations in a knowledge graph for relationships. Thus, concept embedding is expected to truly reflect features of ontological concepts in the embedding space.

3. SYSTEM ARCHITECTURE

The proposed system aims to achieve a better understanding of Bilingual Short text Similarity by making use of two knowledge bases Probase and BabelNet for English, Hindi short text respectively. At the same time employing two different embedding models that is Word Embedding for the words from the short text and Concept Embedding for related concepts of terms from the short text. The similarity of bilingual short texts is computed using the word and concept embeddings obtained from the proposed model.

Figure 1 describes the overall system architecture of the proposed model. First, the Short Texts (ST) are preprocessed by tokenizing and stemming methods. English and Hindi short texts are preprocessed separately. Different stemmers are used for English and Hindi Short text. The preprocessed short text terms are conceptualized by mapping with knowledge bases Probase (English ST) and BableNet (Hindi ST). Word to word correlation network is constructed from short text terms. Similarly, concept to concept network is constructed with associated concepts of terms of short text. Skip gram word2vec model is used to train word to word network and skip gram based concept2vec is used for learning concept to concept network. The resulting word and concept embeddings are used for estimating bilingual short text similarity.

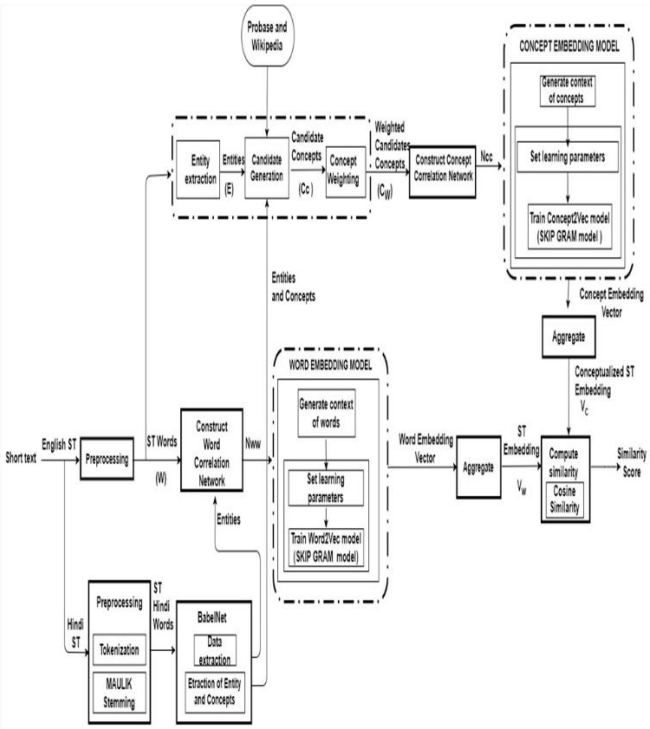


Fig. 1: System architecture

4. EXPERIMENTAL SETUP

4.1 Dataset

Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description.

This report contains a listing of all the sentence pairs comprising the pilot benchmark data set that can be obtained from the link mentioned below:

https://www.researchgate.net/publication/238732681_Pilot_Short_Text_Semantic_Similarity_Benchmark_Data_Set_Full_Listing_and_Description is for evaluating algorithms designed to measure Short Text Semantic Similarity (STSS). A short text is a coherent piece of text at the sentence level, but which does not necessarily conform to the grammatical rules of correctly formed sentences. Thus it includes spoken and typed utterances. This benchmark data set has been used in publications concerning our own STSS measure, now known as STASIS and has been requested by other scientists working in the field. For Hindi Short text we have manually translated every second Short text from all the pairs into Hindi with Romanization of the same available. This makes it possible for even non-Hindi speakers to be able to read the Hindi short text.

4.2 Results

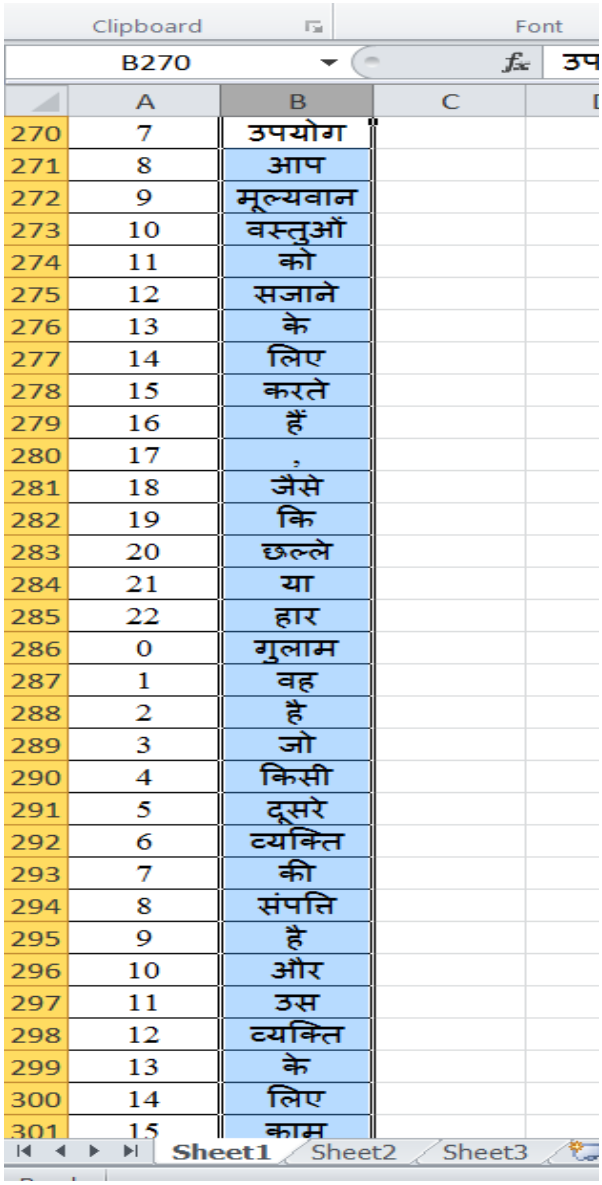
4.2.1 Preprocessing

```
...: df = pd.DataFrame({'col':words})
...: df.to_csv('TokenST.csv', sep=',')
['cord', 'strong', 'thick', 'string', 'smile', 'expression', 'face', 'rooster',
'adult', 'male', 'chicken', 'voyage', 'long', 'journey', 'ship', 'spacecrafe',
'pleased', 'amused', 'friendly', 'noon', 'oclock', 'middle', 'day', 'string', 'thin',
'rope', 'made', 'twisted', 'threads', 'used', 'tying', 'things', 'together', 'tying',
'parcels', 'fruit', 'fruit', 'something', 'grows', 'tree', 'bush', 'contains', 'seeds',
'stone', 'covered', 'substance', 'eat', 'furnace', 'container', 'enclosed', 'space',
'hot', 'fire', 'made', 'example', 'melt', 'metal', 'burn', 'rubbish', 'produce',
'steam', 'autograph', 'signature', 'someone', 'famous', 'specially', 'written', 'fan',
'keep', 'shores', 'shore', 'sea', 'lake', 'wide', 'river', 'land', 'along', 'edge',
'automobile', 'car', 'legends', 'fairy', 'stories', 'wizard', 'man', 'magic', 'powers']

In [9]:
```

Fig. 2: Tokenization (English ST)

Figure 2 shows the result of the tokenization of short texts. The short texts are broken into words, symbols and meaningful elements called Tokens.



	A	B	C
270	7	उपयोग	
271	8	आप	
272	9	मूल्यवान	
273	10	वस्तुओं	
274	11	को	
275	12	सजाने	
276	13	के	
277	14	लिए	
278	15	करते	
279	16	हैं	
280	17	,	
281	18	जैसे	
282	19	कि	
283	20	छल्ले	
284	21	या	
285	22	हार	
286	0	गुलाम	
287	1	वह	
288	2	है	
289	3	जो	
290	4	किसी	
291	5	दूसरे	
292	6	व्यक्ति	
293	7	की	
294	8	संपत्ति	
295	9	है	
296	10	और	
297	11	उस	
298	12	व्यक्ति	
299	13	के	
300	14	लिए	
301	15	काम	

Fig. 3: Tokenization (Hindi ST)

Figure 3 similar to figure 2 shows the result of tokenization of Hindi short texts, where the short text is broken into meaningful elements as Hindi tokens.

4.2.2 Stemming

```
...: df.to_csv('StemST.csv', sep=',')
[',', 'col', '0', ',', 'cord', '1', ',', 'strong', '2', ',', 'thick', '3', ',',
'string', '4', ',', 'smile', '5', ',', 'express', '6', ',', 'face', '7', ',',
'rooster', '8', ',', 'adult', '9', ',', 'male', '10', ',', 'chicken', '11', ',',
'voyag', '12', ',', 'long', '13', ',', 'journey', '14', ',', 'ship', '15', ',',
'spacecraf', '16', ',', 'pleas', '17', ',', 'amus', '18', ',', 'friendli', '19',
'noon', '20', ',', 'oclock', '21', ',', 'middl', '22', ',', 'day', '23', ',', 's
'24', ',', 'thin', '25', ',', 'rope', '26', ',', 'made', '27', ',', 'twist', '28
'thread', '29', ',', 'use', '30', ',', 'tie', '31', ',', 'thing', '32', ',', 'to
'33', ',', 'tie', '34', ',', 'parcel', '35', ',', 'fruit', '36', ',', 'fruit', '
',', 'someth', '38', ',', 'grow', '39', ',', 'tree', '40', ',', 'bush', '41', ',
'contain', '42', ',', 'seed', '43', ',', 'stone', '44', ',', 'cover', '45', ',',
'substanc', '46', ',', 'eat', '47', ',', 'furnac', '48', ',', 'contain', '49', '
'enclos', '50', ',', 'space', '51', ',', 'hot', '52', ',', 'fire', '53', ',', 'm
'54', ',', 'exampl', '55', ',', 'melt', '56', ',', 'metal', '57', ',', 'burn', '
',', 'rubbish', '59', ',', 'produc', '60', ',', 'steam', '61', ',', 'autograph',
',', 'signatur', '63', ',', 'someon', '64', ',', 'famou', '65', ',', 'special',
',', 'written', '67', ',', 'fan', '68', ',', 'keep', '69', ',', 'shore', '70', '
'shore', '71', ',', 'sea', '72', ',', 'lake', '73', ',', 'wide', '74', ',', 'riv
'75', ',', 'land', '76', ',', 'along', '77', ',', 'edg', '78', ',', 'automobil',
',', 'car', '80', ',', 'legend', '81', ',', 'fairi', '82', ',', 'stori', '83', '
'wizard', '84', ',', 'man', '85', ',', 'magic', '86', ',', 'power']
```

Fig. 4: Stemming (English tokens)

After tokenization the tokens of the short text are further preprocessed by porter stemming algorithm for English tokens which removes common morphological endings for words in English and normalizes the texts(figure 4).

	A	B	C
1	जगह		
2	या		
3	स्थिति		
4	को		
5	पागलखाने		
6	के		
7	रूप		
8	में		
9	वर्णित		
10	करते		
11	हैं		
12	,		
13	तो		
14	इसका		
15	मतलब		
16	है		
17	कि		
18	यह		
19	भ्रम		

Fig. 5: Stemming (English tokens)

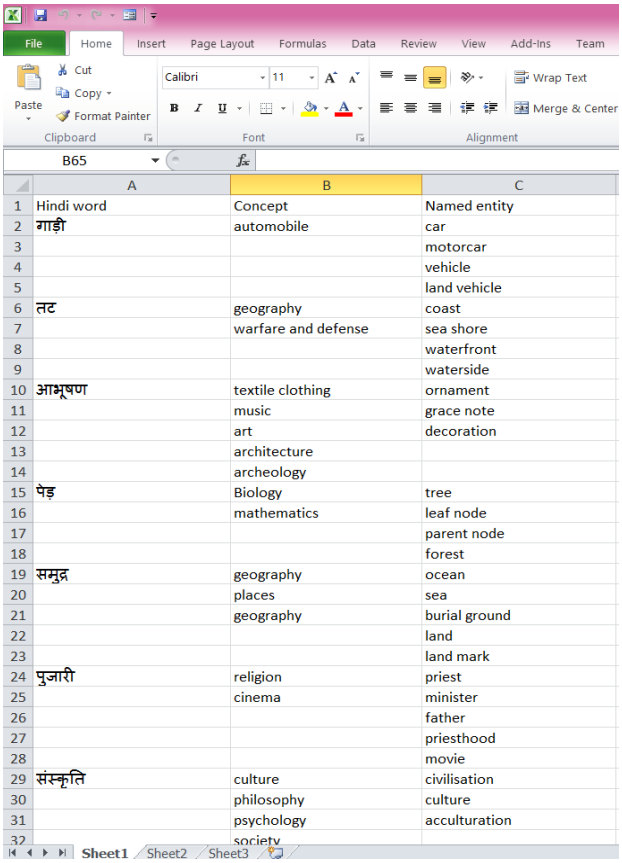
Similar to figure 4, figure 5 shows the result of stemming for Hindi tokens by Maulik stemmer which removes the common morphological endings from words Hindi and normalizes the text.

```
...: doc = nlp('cord strong thick string smile express face rooster adult male
chicken  apple voyage long journey ship America spacecraft pleas amus friend noon
oclock middle day string thin rope made twist thread use tie thing togethe tie parcel
fruit fruits some grow tree bush contain seed stone cover substance eat furnace contain
enclose space hot fire made example melt metal burn rubbish produc steam autograph
signature someon famou special written fan keep shore shore sea lake wide river land
along edg automobil car legend fairy stori wizard man magic power')
...: print([(X.text, X.label_) for X in doc.ents])
[('America', 'GPE'), ('noon', 'TIME'), ('middle day', 'DATE')]

In [13]:
```

Fig. 6: Entity extraction

The result of entity extraction from the stemmed English tokens is shown in figure 6, which is a process of identifying and classifying key elements from text into pre-defined categories.



	A	B	C
1	Hindi word	Concept	Named entity
2	गाड़ी	automobile	car
3			motorcar
4			vehicle
5			land vehicle
6	तट	geography	coast
7		warfare and defense	sea shore
8			waterfront
9			waterside
10	आभूषण	textile clothing	ornament
11		music	grace note
12		art	decoration
13		architecture	
14		archeology	
15	पेड़	Biology	tree
16		mathematics	leaf node
17			parent node
18			forest
19	समुद्र	geography	ocean
20		places	sea
21		geography	burial ground
22			land
23			land mark
24	पुजारी	religion	priest
25		cinema	minister
26			father
27			priesthood
28			movie
29	संस्कृति	culture	civilisation
30		philosophy	culture
31		psychology	acculturation
32		society	

Fig. 7: Hindi concept and entities

Figure 7 shows entity extraction for stemmed Hindi tokens which obtained after the process of identifying and classifying key elements from text into pre-defined categories.

,'pleas 0	[0.02185792349726776]
amus 0	[0.02185792349726776]
friend', 0	[0.005464480874316944]
'noon 0	[0.07650273224043716]
oclock',' 0	[0.005464480874316944]
middle 1	[0.005464480874316944]
day',' 0	[0.005464480874316944]
string 0	[0.005464480874316944]
thin 1	[0.02185792349726776]
rope 1	[0.005464480874316944]
made 1	[0.07650273224043716]
twist', 0	[0.07650273224043716]
'thread 0	[0.005464480874316944]
use 0	[0.005464480874316944]
tie 0	[0.01639344262295082]
thing 0	[0.07650273224043716]
together 1	[0.005464480874316944]
tie', 0	[0.060109289617486336]
'parcel 0	[0.01639344262295082]
fruit','fruits 0	[0.07650273224043716]
some 0	[0.005464480874316944]
grow 0	[0.005464480874316944]
tree', 0	[0.060109289617486336]
'bush 0	[0.005464480874316944]
contain 0	[0.060109289617486336]
seed', 0	[0.010928961748633888]
'stone 0	[0.02185792349726776]
cover 0	[0.005464480874316944]
substance','eat 0	[0.005464480874316944]
furnace 1	[0.005464480874316944]
contain 0	[0.005464480874316944]
enclose 0	[0.07650273224043716]

Fig. 8: 7 IDF of entities

Figure 8 IDF of Entities For the extracted English entities weights are assigned by the means of IDF is shown in figure 8.

Entity	Concept	prob
America	developed country	0.228
America	western country	0.104
America	economy	0.076
America	place	0.065
America	area	0.053
America	jurisdiction	0.080
America	market	0.105
midday	time	0.137
midday	offpeakhour	0.129
midday	peakhour	0.137
midday	agreed time	0.018
midday	photography	0.069
midday	classic movie	0.136

Fig. 9: Mapping Entity to Concepts in Probase

The entities extracted from the previous module is mapped with their related concepts in the knowledge bases probase and Wikipedia (figure 9).

	Entity	Concept	prob	Rank_by_prob
10	midday	agreed time	0.018	1.0
15	noon	joyce scholar	0.040	2.0
4	America	area	0.053	3.0
16	noon	audit study	0.054	4.5
14	noon	usage hour	0.054	4.5
3	America	place	0.065	6.0
11	midday	photography	0.069	7.0
2	America	economy	0.076	8.0
5	America	jurisdiction	0.080	9.0
1	America	western country	0.104	10.0
6	America	market	0.105	11.0
8	midday	offpeakhour	0.129	12.0
12	midday	classic movie	0.136	13.0
7	midday	time	0.137	14.5
9	midday	peakhour	0.137	14.5
13	noon	hour	0.161	16.0
0	America	developed country	0.228	17.0

Fig. 10: Concept Selection

From the mapped and weighted entities concepts are generated and typicality score is calculated (figure 10).

top entity (character)	concept (character)	prob (double)
America	developed country	0.228
midday	time	0.137
noon	hour	0.161
bird	animal	0.421
Jeweley	item	0.294
Rooster	animal	0.206
tree	object	0.209
fruit	food	0.439

Fig. 11: Candidate concept

Based on the typicality score the top k candidate concepts are selected and concept cleaning is done to remove redundant concepts (figure 11).

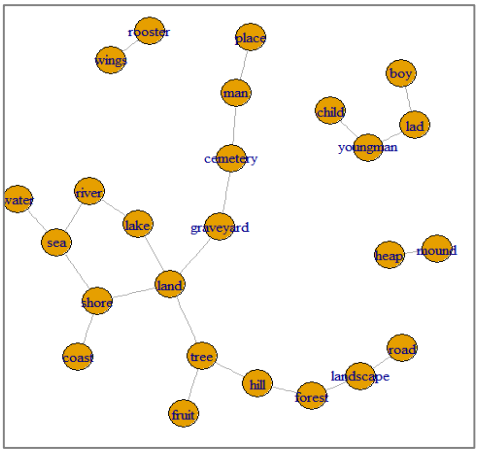


Fig. 12: Word correlation network

Figure 12 shows the word to word correlation network constructed with (word, word) pair to establish all the possible relationship of a word with other related words.

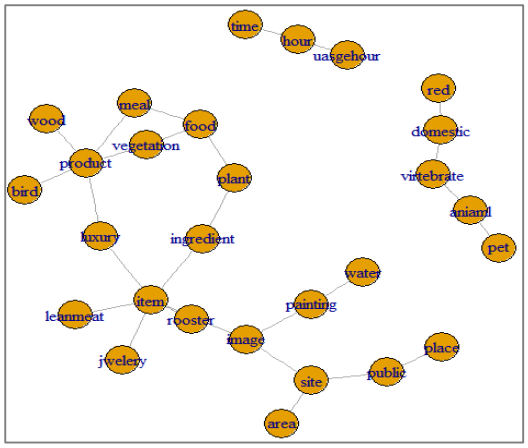


Fig. 13: Concept correlation network

Figure 13 shows the concept to concept correlation network constructed with (concept, concept) pair to establish all the possible relationship of a concept with other related concepts.

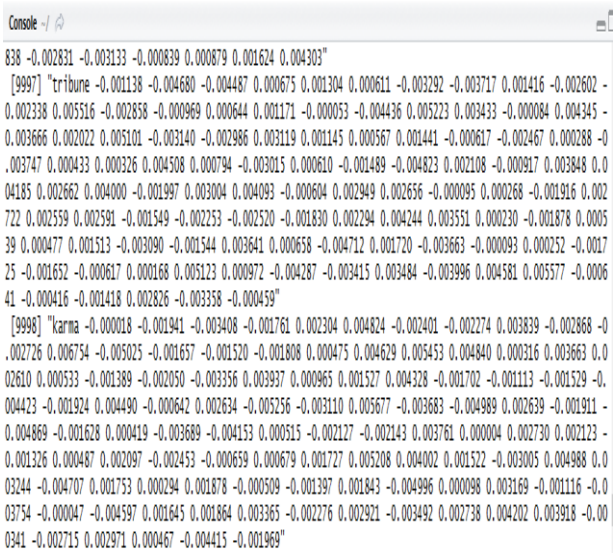


Fig. 14: Word embedding

Figure 14 shows the word vectors obtained from the word embedding model (word2vec), where every word is converted into corresponding vector values.

Fig. 15: Concept embedding

The concepts are also converted into corresponding concept vectors using the concept embedding model (concept2vec) the concept vector values are shown in figure 15.

4.3 Evaluation metrics

4.3.1 Performance measures used for conceptualization

Precision

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \quad (1)$$

Recall

$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ negative} \quad (2)$$

F-measure

$$\text{F measure} = 2 * \frac{Presicion * Recall}{Presicion + Recall} \quad (3)$$

4.3.2 Short text

Bilingual short text pair

ST pair 1: Fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.

एक भट्ठी एक कंटेनर या संलग्न स्थान है जिसमें बहुत गर्म आग बनाई जाती है, उदाहरण के लिए धातु पिघलाने के लिए, जला हुआ कचरा या भाप का उत्पादन करें।

ST pair 2: The shores or shore of a sea, lake or wide river is the land along the edge of it.

वुडलैंड बहुत सारे पेड़ों के साथ भूमि है।

ST pair 3: A cemetery is a place where dead people's bodies or their ashes are buried.

एक कब्रिस्तान भूमि का एक क्षेत्र है, कभी-कभी एक चर्च के पास, जहां मृत लोगों को दफनाया जाता है।

Table 1: Conceptualization evaluation

Conceptualization		Precision	Recall	F measure
Without probase	ST pair 1	0.71	0.86	0,777
	ST pair 2	0.45	0.47	0.459
	ST pair 3	0.73	0.72	0.724
With probase	ST pair 1	0.80	0.87	0.833
	ST pair 2	0.55	0.60	0.573
	ST pair 3	0.78	0,85	0.813

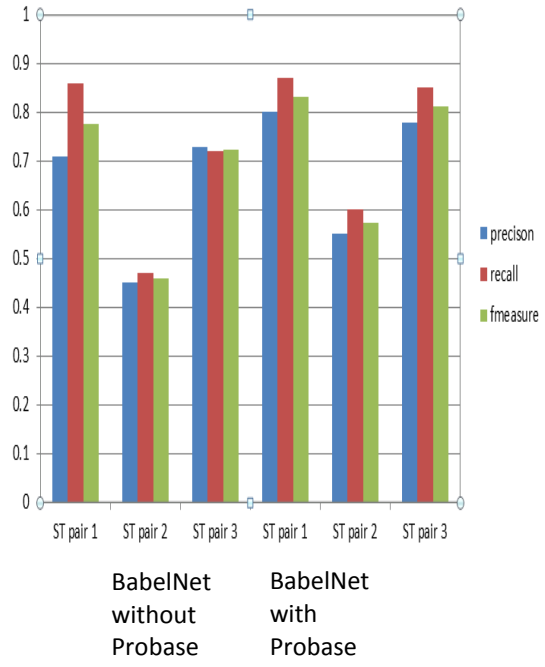


Fig. 16: Conceptualization evaluation

The graph in figure 16 shows results for evaluation of conceptualization for the proposed framework. We can observe that improved results are obtained for conceptualization with Probase compared to conceptualization without Probase.

Table 2: Short text Similarity Evaluation

	Word embedding Score	Concept embedding Score
Correlation of similarity with Pilot dataset similarity score	0.9580506	0.9755498

The correlation of similarity was computed for all the 65 short text pairs in the pilot short text dataset with the scores of similarity obtained by word embedding and concept embedding for the same 64 short text pairs.

4.4 Test cases

4.4.1 ST pair 1: An automobile is a car.

In legends and fairy stories, a wizard is a man who has magic

किंवदंतियों और परियों की कहानियों में, एक जादूगर एक ऐसा व्यक्ति हैजिसके पास जादुई शक्तियां हैं।

4.4.2 ST pair 2: The coast is an area of land that is next to the sea.

A hill is an area of land that is higher than the land that surrounds it.

हाड़ी भूमि का एक क्षेत्र है जो उस भूमि से अधिक है जो इसे घेरती है।

4.4.3 ST pair 3: A magician is a person who entertains people by doing magic tricks.

In legends and fairy stories, a wizard is a man who has magic powers

किंवदंतियों और परियों की कहानियों में, एक जादूगर एक ऐसा व्यक्ति है जिसके पास जादुई शक्तियां हैं|

Table 3: Test cases

S no.	ST pair	Pilot score	Word embedding	Concept embedding
1.	I	0.26	0.15	0.19
2.	II	0.53	0.38	0.49
3.	III	0.92	0.85	0.89

5. CONCLUSION

The proposed system has optimized the way of computing similarity for Bilingual short text. Mapping with the BabelNet concepts and entities helps us to retrieve the wide range of related concepts of Hindi terms. Conceptualization with Probase ensures higher precision, recall and F measure. The learning of word and concept correlation networks improves the semantic relationship between words and concepts of short text the fact that making use of two different embedding models (Word2vec, Concept2vec) gives a better correlation of bilingual short text similarity score.

6. FUTURE WORK

In future, the proposed framework can be extended further to support Multilingual Short text Embedding that would present a novel technique for creating more efficient Natural Language Processing models, which will be much faster, simpler and scalable. It would adapt to new languages and new tasks while achieving strong results across many languages.

7. REFERENCES

[1] W. Wu, H. Li, H. Wang, and K. Q. Zhu (2012) "Probate: A probabilistic taxonomy for text understanding," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 481–492.

[2] Natasha J, Karpagapriya K, Vijayarani J (2018) "Conceptualized Short Text Embedding Using Knowledge Base" in Journal of Emerging Technologies and Innovative Research (JETIR) International Open Access Journal, ISSN2349-5162. https://www.google.com/search?source=hp&ei=7ppGXMWfLtmWwgP62o_4BA&q=http%3A%2F%2Fwww.jetir.org%2Fview%3Fpaper%3DJETIR1812600&btnK=Google+Search&oq=http%3A%2F%2Fwww.jetir.org%2Fview%3Fpaper%3DJETIR1812600&gs_l=psy-ab.3...920.920..1462...0.0..0.136.241.0j2.....0....2j1..gws-wiz.....0.93Aov0z84QQ

[3] Peipei Li, Haixun Wang, Kenny Q. Zhu, Zhongyuan Wang, Xindong Wu (2013) "Computing Term Similarity by Large Probabilistic isA Knowledge".

[4] Ta-Chung Chi, Ching-Yen Shih, Yun-NungChen (2018) "BCWS: Bilingual Contextual Word Similari ty", National Taiwan University, Taipei Taiwan.

[5] Goran Glavaš , Marc Franco-Salvador , Simone P. Ponzetto , Paolo Rosso (2017) "A resource-light method for cross-lingual semantic textual similarity Goran" ELSEVIER-Knowledge Based System.

[6] Kerstin Denecke (2008) "Using SentiWordNet for Multilingual Sentiment Analysis" Research CenterL3S Appelstrasse 9a, D-30167 Hannover, Germany.denecke@l3s.de

[7] Edi Faisal, Farza Nurifan, Riyanarto Sarno (2018)"Word Sense Disambiguation in Bahasa Indonesia Using SVM" 2018 International Seminar on Application for Technology of Information and Communication (iSemantic).

[8] RobertoNavigli, SimonePaoloPonzetto (2010) "BabelNet: Buildingal Very Large Multilingual Semantic Network" Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 216–225, Uppsala, Sweden, 11-16 July 2010 Association for Computational Linguistics.

[9] Faisal Alshargi, Saeedeh shekkarpour, Tommaso Soru, Amit Shet (2019) "Concept2 vec:Metrics for Evaluating Quality of Embeddings for Ontological Concepts" University of Leipzig, Germany.

[10] Yuan Ni, Qiong Kai Xu,Feng Cao Yosi, Haifa yosimass, Haifa dafna, Hui Jia Zhu zhuhuij Sheng Ca (20 16)"Semantic Documents Relatedness using Concept Graph Representation" ACM.

[11] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, Xiaofang Zhou (2015) "Short Text Understanding Through Lexical-Semantic Analysis" ICDE Conference

[12] Haoyu Pu, Gaolei Fei, Hailin Zhao, Guangmin Hu Chengbo Jiao (2017) "Short Text Similarity Calculation Using Semantic Information" 2017 3rd International Conference on Big Data Computing and Communications

[13] Amir H. Jadidinejad(2013) "Unsupervised Information Extraction using BabelNet and DBpedia" Islamic Azad University,

