Volume 8, No. 5, May-June 2017

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Hybrid Approach in Classification of Telugu Sentences

G.Pratibha
Asst.Professor, CSE
Matrusri Engineering College, Hyderabad
Telangana, India

Dr.Nagaratna P Hegde
Professor, CSE
Vasavi College of Engineering & Technology, Hyderabad
Telangana, India

Abstract: Natural Language Processing (NLP) is a multidisciplinary research area that explores how computer understand human language in the form of text or speech to do useful things. Telugu is most prominent and morphologically rich dravidian language spoken by around one million people. When research in computational linguistics is concerns telugu is far behind other south Indian languages .Recognizing sentence similarity is most useful task in all the languages which are useful for improving plagiarism detection of documents, word sense disambiguation, query evaluation, paraphrase detection and question answering. In this paper, we discuss an hybrid approach to calculate semantic similarity score between two telugu sentences using supervised learning for classifying simple telugu verb less sentences using linguistic knowledge of telugu language and combination of rule based and stochastic methods are used to measure similarity between sentences.

Keywords: NLP, supervised learning, Rule based , stochastic , Linguistic knowledge , Similarity Measures.

1. INTRODUCTION

There are several methods to measure sentence similarity they are all well suited for English language but, when we implement them for telugu, They are giving poor accuracy. The reason is telugu words are highly morpho_inflected with GNP (gender, number,person) .Tense, aspect and modality i.e. TAM and they will play an important role in syntactic and semantic representation of telugu language sentences .Most of the similarity methods evaluate using BOW, skip gram and skip thought processes, which can give high level of syntax level similarity but achieved low level of semantic similarity. If we compare the following sentences by any of the existing method which gives them as more similar words but, they are completely different.

WX: nA BArya reVMdo wammudu.
My wife second brother.

WX: nA reVMdo wammudi BArya.
My second brother's wife

All the telugu sentences may not behave same .First of all, they will be classified into different groups according to their behavior [1]
In this paper, we classified the simple sentences based on the deep linguistic knowledge [1].
Telugu sentences are classified into simple, complex and coordinate sentences [1,2]. Simple sentences are independent sentences. Complex sentences are formed by inserting a sentence into a determined position of other sentence. Coordinate sentence is formed by joining a sentence to another sentence. The original sentences gets modified when we join one sentence to others. The resultant sentence becomes complex and coordinate sentences. If we observe
them keenly, then there is no doubt that they have fundamental rule of simple sentences only. Hence first we analyze the simple sentences. First of all simple sentences are primarily divided into two categories as

i. KriyA rahitha (without verb)
ii. KriyA sahitha (with verb)
A new model is proposed to calculate semantic similarity for kriyA rahitha (without verb) simple sentences are proposed[1].
Algorithm of this model has two phases.
i) Classification phase.
ii) Similarity phase.

2. PROPOSED WORK

Algorithm:
Input: telugu_UTF text with verb less simple sentences.
Output : Semantic similarity score.
Method:
Classification Phase:
i) Split the text into sentences.
ii) Convert them into WX notation.
iii) Label each sentence with a suitable clause label as 1 to 11.
iv) Convert it into Vector using Sentence2Vec method.
v) Split the data into Training and Test data sets using cross fold validation.
vi) Train the Classifier with training data set.
vii) Classify the test data using various Machine learning tools.
Similarity Phase: It is also called as similarity phase where we compare similarity between sentences.Phase1
classifies the sentences into the classes among 1-11.Read them for further calculating the semantic similarity of sentences. Here we use different hybrid similarity methods such as combination of rule based and stochastic measures for different classes of sentences. In this paper Classification phase of Algorithm for simple verb less sentences are explained.
Classification Phase: This phase is explained with the following figure Fig. 1.

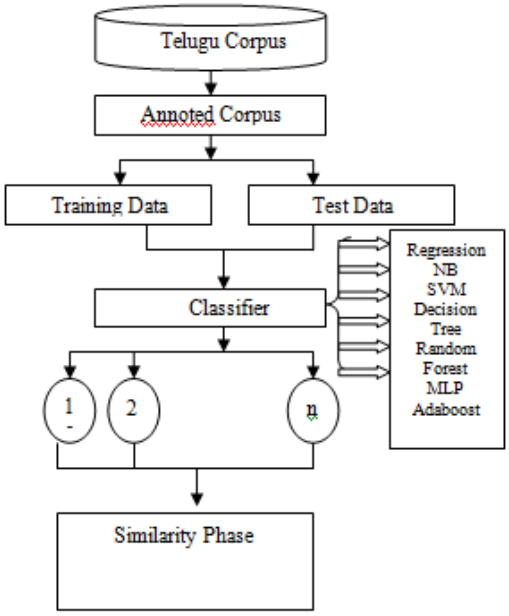

Fig. 1:Classification Phase

3. IMPLEMENTATION

KriyA rahitha Simple sentence : It is a simple sentence without any verb. Some of these sentences will be naturally without verb. In these sentences usually two noun phrases(NP) will be there. In that one is subject (can be a noun or pronoun)and another one is object(noun usually depends of adjectives).

Preprocessing: Dataset is prepared with verb less telugu simple sentences and it is domain independent. Split the text into sentences then convert them into wx form [4,7]. These data is fed to Doc2Vec tool Provided by a python module Gensim to generate corresponding feature vector file with hundred features for every sentence of input.Doc2Vec gives the distributed semantic representation of given sentence.

Annotated Data: In this step each sentence is attached with a class label number in the range of 1 to 11.Telugu sentences are classified into simple, complex and compound sentences. Simple sentences are again simple with verb and without verb. In this paper we deal with semantic similarity between simple sentences without verb. They have been divided into eleven classes depending upon their contextual representation [1].

Classification: In order to classify a sentence .The classifier must be trained with training data. vector file has split into training and test data sets applying 5-fold cross validation. These classifier models trained with training data and labels.

Classifiers:

There are several classifiers to predict our test data from trained data. These classifiers are used from python module scikit-learn [3], which has all the built-in classifiers. We used various classifiers[5,6,8] to classify our telugu data.

Regression: Liner regression is one powerful classifier, which is poly class logistic model. It is used to classify an object in the predefined classes by using probability with the help of independent variables.

Naïve Bayes: Naïve bayes is also multi-class probability classifier uses Bayes Theorem. It predicts the class for an object based on conditional probability. It estimates the class labels based on previous occurrence. This algorithm works efficiently for the functions which are linearly separable and also reasonably good for linearly inseparable.

Support Vector Machine: It is a non probabilistic model and works similar to neural networks**.** This classifier works with supervised algorithm and it generates many hyper planes in a high dimensional vector space to classify the objects.

Neural Network: Multi-Layer Perceptron is a classifier which classifies an objects based on neural network. It is a multi layer feed forward neural network. It has one input layer, one output layer and multiple hidden layers. Input layer is with neurons equals to number of inputs, at which we feed our input. It has one output layer has neurons equals to number of classes which we have. Hidden layers are all intermediate layers from input to output which are all connected.

Decision Tree: This classifier classifies with a decision based model. It is a tree like classification in witch each internal node is labeled with a feature and leaf is labeled with class label.

Random Forest: This classifier is an ensemble method of decision trees. It is hierarchical decision tree method which constructs multiple decision trees and calculates score from each of them and finally all will be compared to get final score.

Adaboost Ensemble: This classifier is used to enhance the weak classifiers on repeatedly enhanced versions of data. The classes are predicted by taking average of all the iterations.

Class1: In this class, There are two noun phrases. First one acts as subject can be a noun or pronoun and second is adjective dependent which plays key roll in semantic similarity computation.with other sentence by mapping the derived suffix after stemming the word from POS tagger.

maMcivAru → maMci+vAru.

Now the dependent phrase vixyArwulu is compared and get matched(plural) with suffix vAru. Usally suffixes will be vAru,vAdu,xi,vALLu

Wx: A vixyArWulu maMci vixyArWulu
Those students are good students.

Wx: A vixyArWulu maMcivAru.
Those students are good.

Class2: In this class, we compare noun phrases of first sentence with corresponding NPs in the second sentence with help of synset in the database.

Wx:Ayna vyApAri.
He Businessman

Class3 :In this class, base of head words of two sentences are same (pos tagger) and suffix is “lamu” is matched from the rule of GNP(gender,number and person) of PRP (nenu) singular and (memu)plural.

Wx: nenu vixyArwi
I am student.
Wx: memu vixyArwulamu.
We are students.

Class4: Sentences will also be similar even though the words exchange their positions. Cosine similarity is used here to classify the sentences.

Wx: vAlylyu amAyakulu
Those are innocent.

Class5: Sentences can also be similar when the words exchanges their positions with a little variations in the morphological inflectional. Here word based similarities are used upon words after stemming by taking headwords into consideration.
Wx: vAde vIdu
Wx:vIde vAdu
Class6:The below sentences are ending with adjective wich are not measurable. sometimes that adjective changes its position .Similarities are by considering synset of the WordNet.
WX: I ammAyi welupu.
This girl fair.
WX: Koyila nalupu.
Kukoo black.
WX: I abbAyi poVtti.
This boy short.
Class7:In this we classify sentences as ends with measurable adjective.We measure the similarity for these sentences by wordbased similarity.
Wx: A koVMda eVwwu 3000 adugulu.
That mountain height 3000 ft.
Wx: A koVMda 3000 adugula eVwwu.
That mountain 3000 ft height.
Class8: In this case we handle comparative sentences which are impossible to measure with cosine measure which is 0.0 . POS tagging also not suitable. This case is treated as an exception and handled with rule based method by considering database of antonyms.
Wx: awanu nA kaMTe peVxxa.
He I than elder
Wx: awani kanna nenu cinna.
He than I younger.

Class9: In this , we handle relative sentences when –ki/ku case marker to the person and similarity is measured using cosine similarity as 1.0.
Wx: awanu nAku wammudu.
He me brother.
Class10: In this, we handle verb less sentences which are dealing with tense. In this we compare words by considering their tenses in the database .
Wx: awanu upAxyAyudu.
He teacher.
Wx: axanu upAxyAyudu ayyAdu.
Wx: axanu upAxyAyudu avuxAdu.
In the above sentences when tense based words are replaced by negative words like kAdu, kAmu, kAdhu, kAvu,
kAru, kAnu then they will be labeled as contradicts and they are dissimilar.
Case11: There is one more clause of sentences which will be similar by word based similarity and semantically meaning less if we change their position of adjective before noun, but, we can make them similar by taking NP into consideration. This is handled with the help of pos tagger.
Wx: I pATaM kaRtaM.
This lesson hard.

4. RESULTS

Results are taken from four iterations and average is taken from all. Some of the classes are giving low accuracy due to classifiers are trained with less data for that class.
We can improve the accuracy by increasing the training data set size.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
LRC	77.02	88.91	85.21	85.43	87.03	82.11	77.02	88.91	85.21	85.43	87.03
NBC	76.03	66.11	67.56	82.43	81.01	65.17	76.03	66.11	67.56	82.43	81.01
SVM	71.32	69.42	67.32	62.45	72.45	66.65	71.32	69.42	67.32	62.45	72.45
MLP	72.92	79.22	67.57	66.54	71.45	69.53	72.92	79.22	67.57	66.54	71.45
DT	66.98	61.35	66.01	63.98	56.54	58.93	66.98	61.35	66.01	63.98	56.54
RF	72.45	82.65	81.61	82.56	76.43	72.63	72.45	82.65	81.61	82.56	76.43

5. CONCLUSION

Telugu is semantically rich language. Developing a similarity detection system is not an easy task as English language. It is not possible by purely statistical tools .Hence, I made an attempt to improve the accuracy by developing rule based classifiers with deep linguistic knowledge.

6. FUTURE WORK

In this paper, Classification is done for verb less simple sentences. It is also required for all other simple sentences, complex, Compound and a special case of imitative sentences.

7. REFERENCES

[1] Chekuri Rama Rao, “Telugu Vakyam”, Andhra Pradesh Sahithya Academy, 1975.
[2] Bh.Krishna Murthy and J.P.L.Gwynn, A Grammar of Modern Telugu,. Oxford University press, 1985.
[3] “Scikit-learn: A machine learning in python”, http://www.Scikit-learn.org.
[4] Quoc V. Le and Tomas Mikolov, ”Distributed Representation of Sentences and Documents.”, arXiv preprints, arXiv,1405.4053, 2014.
[5] Martin Anthony and Peter L. Bhartlett, “Neural Network Learning: Theoretical Foundations”, Cambridge University Press, 1999.
[6] Ng, Andrew Y. and Jordan, Michael I, “On Discriminative Vs. Generarative Classifiers : A Comparison of logistic regression and naïve Bayes”, Advances in Neural Information Processing Systems 14,NIPS 2001.
[7] Zellig S Harris, Distributional Structure, Word. 10-23 : 146-162, 1954.
[8] Keerthi, S.S, Shavade, S. K, Bhattacharya. C.,& Murthy, K.R.K. “Improvements to Platt’s SMO algorithm for SVM Classifier.”, Neural Computation, vol. 13, Issue 3, March 2001, pp. 631-649.
[9] Jiang, J .J. Cornath, D.W, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, arXiv preprint, arXiv: cmp-lg /9709008, pp. 1-15,1997.