Final Mini Project Demonstration

UE19CS306D

Project Title : Feature Extraction for Paraphrase Detection in a Low Resourced Language : Kannada

Project Guide : Prof. Mamatha H R

Project Team : Anagha H M (PES1UG19CS057)

Karthik Sairam (PES1UG19CS210)

Project Abstract and Scope

Paraphrase detection is a Natural Language Processing classification problem. The main goal of paraphrase identification is to find if two phrases of different lengths are synonymous. This is done by finding the similarity between the given two paraphrases. This paper attempts to preprocess and extract the necessary features required to perform paraphrase detection in Kannada. This is the first few steps done before proceeding to paraphrase identification.

Literature Survey

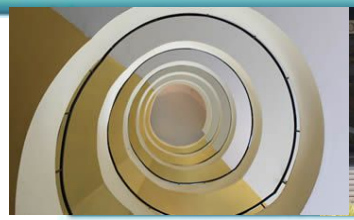
	Paper 1	Paper 2
Title	A Survey on Paraphrase Detection Techniques for Indian Regional Languages	Detection of paraphrases for Devanagari languages using support vector machine
Authors	Srivastava, Shruti & Govilkar, Sharvari	D. S. Bhole and S. S. Patil
Year of Publishing	2017	2018
Inferences	Explains the approach of NLP in English, and provides information on different Paraphrase detection techniques and Similarity Metrics used in NLP.	Details about the feature extraction po of paraphrase detection, including Tokenization, Stop-word elimination, Stemming
Shortcomings	The scope is limited to English only, hence significant changes have to be made to implement a model in a low resourced language like Kannada.	Done for Hindi and Marathi, and since Kannada is significantly grammatically complicated than hindi, a more robust model is required to classify paraphrases.

Literature Survey

	Paper 3	Paper 4
Title	Paraphrase plagiarism identification with character-level features	Paraphrase Detection for Short Answer Scoring
Authors	Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gómez, M	Koleva, N., Andrea Horbach, Alexis Palmer, Simon Ostermann and Manfred Pinkal.
Year of Publishing	2019	2014
Inferences	Proposed a model that extracts 6 features for paraphrase detection. Also proposed method for corpus creation.	Two methods - Basic and Chunk phrased ; uni and bi directional. Also extracts 5 features.
Shortcomings	Talks only about specific application implementation.	Done in German.

Literature Survey

	Paper 5	Paper 6
Title	A survey on paraphrase recognition.	Hindi Paraphrase Detection using Natural Language Processing Techniques & Semantic Similarity Computations
Authors	Magnolini, Simone	Kanjirangat, Vani & Gupta, Deepa
Year of Publishing	2014	2016
Inferences	Defines Paraphrasing and explains different algorithms/approaches. Comparison between different approaches is given	Tagging, pruning, stop word removal and stemming
Shortcomings	Doesn't give the exact procedure for doing paraphrase identification.	-






Literature Survey

	Paper 7	Paper 8
Title	Paraphrase Detection in Hindi Language using Syntactic Features of Phrase	An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine
Authors	Bhargava, Rupal & Baoni, Anushka & Jain, Harshit & Sharma, Yashvardhan	P. Vigneshvaran, E. Jayabalan and A. V. Kathiravan
Year of Publishing	2016	2014
Inferences	Stemming and Soundex core used. Error analysis and classifier comparison already done	The approach specified here is Tokenization, POS Tagging, Token Match, Token Count. The features extracted are thus fed to the SVM.
Shortcomings	We can inherently use the best classifier but this is done only for hindi	-

Literature Survey

	Paper 9	Paper 10
Title	Paraphrase Detection Using Manhattan's Recurrent Neural Networks and Long Short-Term Memory	A Paraphrase and Semantic Similarity Detection System for User Generated Short-Text Content on Microblogs
Authors	A. A. Aziz, E. C. Diamal and R. Ilyas	Kuntal Dey, Ritvik Shrivastava, Saroj Kaushik
Year of Publishing	2019	2016
Inferences	Converts word to vector. Creates a matrix. Manhattan LSTM offers an approach to common sentence similarity problems. Uses a Siamese Neural Network.	Topic phrase is removed, normalization is performed, boundary correction is done , cleaning and synonym, hypernym are performed
Shortcomings	Extensive process. Training data is huge.	Used traditional classifier. Deep learning would have enabled usage of more features.

Project Requirements

SL. No	Functional Requirements	Non-functional Requirements
1	Create an appropriate dataset	Clean the sentence pools
2	Input texts in Kannada	Extract the features from the cleaned data

Methodology

1. Creation of the dataset

A
Sentence_1
ಮಾಲ್ ದಾಳಿಕೋರರು 'ಕಡಿಮೆ ಹೆಚ್ಚು' ತಂತ್ರವನ್ನು ಬಳಸಿದ್ದಾರೆ
ಕಾಂಬೋಡಿಯಾದ ಕಾರ್ಮಿಕರ ಸಾವಿನ ಸ್ಮರಣಾರ್ಥ ವಿರೋಧ ಪಕ್ಷದ ನಾಯಕರು ಹೊರಹೊಮ್ಮುತ್ತಾರೆ
ದುರ್ಬಲ ಗಳಿಕೆಯು ಷೇರುಗಳನ್ನು ಕಡಿಮೆಗೊಳಿಸುತ್ತದೆ
ಸೊಲೊಮನ್ ದ್ವೀಪದಲ್ಲಿ 6.8 ಭೂಕಂಪ ಸಂಭವಿಸಿದೆ
ಶಾಂಘೈನಲ್ಲಿ ದ್ರವ ಅಮೋನಿಯಾ ಸೋರಿಕೆ 15 ಸಾವು
ಸಿರಿಯಾದ ಸ್ನೇಹಿತರು ವಿರೋಧ ಒಕ್ಕೂಟವನ್ನು ಸಿರಿಯನ್ ಜನರ ಕಾನೂನುಬದ್ಧ ಪ್ರತಿನಿಧಿಯಾಗಿ ಗುರುತಿಸುತ್ತಾರೆ
ಭಾರತೀಯ ಮಾಧ್ಯಮ: ಕಾಮನ್‌ವೆಲ್ತ್ ಶೃಂಗಸಭೆ
ರಷ್ಯಾ ವಿಮಾನ ದುರಂತದಲ್ಲಿ ಸಾವಿನ ಸಂಖ್ಯೆ ಏರಿಕೆಯಾಗಿದೆ
ಮನೆ ತಂಗಿ' ಹಗರಣದಲ್ಲಿ 7 ಮಂದಿ ಬಂಧನ
ಇರಾಕ್‌ನಲ್ಲಿ ಆತ್ಮಹತ್ಯಾ ಬಾಂಬರ್ ಅಂತ್ಯಕ್ರಿಯೆಯ ಮೇಲೆ ದಾಳಿ
ಬೋಯಿಂಗ್ 787 ಡ್ರೀಮ್‌ಲೈನರ್‌ಗೆ ಬೆಂಕಿ ಹತ್ತಿಕೊಂಡಿತು; ಸ್ಯಾಕ್ ಹೊಡೆತವನ್ನು ತೆಗೆದುಕೊಳ್ಳುತ್ತದೆ
ಸಿರಿಯಾ ಮಿಲಿಟರಿ ಪೊಲೀಸ್ ಮುಖ್ಯಸ್ಥರು ವಿರೋಧಕ್ಕೆ ಪಕ್ಷಾಂತರ
ಬೆಲ್ಜಿಯಂ ಕೋಚ್ ಅಪಘಾತದಲ್ಲಿ ಐವರು ಸಾವು
ನೈಜೀರಿಯಾದಲ್ಲಿ ಬಂದೂಕುಧಾರಿಗಳು ಐಳು ವಿದೇಶಿ ಕಾರ್ಮಿಕರನ್ನು ಅಪಹರಿಸಿದ್ದಾರೆ
ಇಸ್ರೇಲ್ ವಿಮಾನಗಳು ಸಿರಿಯಾದೊಳಗೆ ದಾಳಿ ಮಾಡುತ್ತವೆ
ಪಶ್ಚಿಮ ದಂಡೆಯಲ್ಲಿ ಇಸ್ರೇಲಿ ಪಡೆಗಳೊಂದಿಗೆ ಪ್ಯಾಲೆಸ್ತೀನಿಯರು ಘರ್ಷಣೆ ನಡೆಸಿದರು
ಹೊಸ ದರ್ಫರ್ ಬುಡಕಟ್ಟು ಘರ್ಷಣೆಯಲ್ಲಿ 10 ಮಂದಿ ಸಾವನ್ನಪ್ಪಿದರು

Features of the created dataset

1. Finding an appropriate dataset is indeed challenging due to :
   a. Lack of meaningful datasets in the Kannada language
   b. Lack of datasets containing the structure appropriate for our project

2. Hence, a dataset was created for the sole purpose of this project. This contained data points in the Kannada language, and the following features:

   a. Over 600 data points
   b. Over 1200 sentence in Kannada
   c. 2 Columns, each containing a sentence pool. For a given row, the sentences from the 2 pools might/might not be paraphrases of each other (variedness).

2. Cleaning of the dataset

ಭಾರತೀಯ ಮಾಧ್ಯಮ ಕಾಮನ್‌ವೆಲ್ತ್ ಶೃಂಗಸಭೆ
ರಷ್ಯಾ ವಿಮಾನ ದುರಂತದಲ್ಲಿ ಸಾವಿನ ಸಂಖ್ಯೆ ಏರಿಕೆಯಾಗಿದೆ
ಮನೆ ತಂಗಿ ಹಗರಣದಲ್ಲಿ ಮಂದಿ ಬಂಧನ

ನಾಯಕರು ಆಫ್ರಿಕಾ ದಿನಾಚರಣೆಗಾಗಿ ಭೇಟಿಯಾಗುತ್ತಾರೆ
ಉತ್ತರ ಕೊರಿಯಾ ಕೇಸಾಂಗ್ ಪ್ರವೇಶವನ್ನು ನಿರ್ಬಂಧಿಸುತ್ತದೆ

3. Feature Extraction

This involves tokenization and stemming.

1. Tokenization can be done successfully, and is discussed in later slides.
2. Stemming is an integral part of feature extraction, and is discussed later.

ಒಬಾಮಾ ಒಬಾಮಾಕೇರ್‌ಗೆ ಸಹಿ ಹಾಕಿದರು

['_ಒಬಾಮಾ', '_ಒಬಾಮಾ', 'ಕೇ', 'ರ್', '_ಗೆ', '_ಸಹಿ', '_ಹಾಕಿದರು']

Technologies Used

- Python 3.8
  - Pandas
  - Numpy
  - Regex
- INLTK (Indian Natural Language Toolkit)

Research Questions

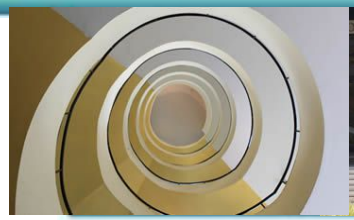
1) Why it is tougher to tokenize the words in Kannada?

- English translation of “ in him “ = ಅವನಲ್ಲಿ
- The observation:
  - The same two words are syntactically written as one word in Kannada
  - We have complex grammar that a language like English, such as :
    - Sandhi
    - Vibhakti Prathyaya
    - Samasa
- We need to implement in such a way so as to handle such special cases of the Kannada language.

Research Questions

2) How can stemming be implemented? Is it necessary?

- Stemming is an integral part of “most” NLP tasks
- If a document exists with > 10,000 words, probability of having tokens from the same root word is high.
- But, paraphrase detection works on just two sentences.
- What is probability of having same stems in those 2 sentences?
  - Very Less and
  - Negligible






Algorithm

A general approach to applying the feature extraction:

1. Begin

2. For every sentence Si in sentence pool S, do

3. Tokenize each word in Si such that tokens are both:
   a. Word-based and
   b. Syntactic-based

4. Repeat Steps 2 and 3 for all sentence pools

5. End

References

[1] Sarkar, Sandip & Saha, Saurav & Bentham, Jereemi & Pakray, Dr. Partha & Gelbukh, Alexander. (2016). NLP-NITMZ@DPIL-FIRE2016:Language Independent Paraphrases Detection.

[2] Srivastava, Shruti & Govilkar, Sharvari. (2017). A Survey on Paraphrase Detection Techniques for Indian Regional Languages. International Journal of Computer Applications. 163. 42-47. 10.5120/ijca2017913757.

[3] Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gómez, M. et al. Paraphrase plagiarism identification with character-level features. Pattern Anal Applic 22, 669–681 (2019). https://doi.org/10.1007/s10044-017-0674-z

[4] Koleva, N., Andrea Horbach, Alexis Palmer, Simon Ostermann and Manfred Pinkal. “Paraphrase Detection for Short Answer Scoring.” (2014).

References

[5] Koleva, N., Andrea Horbach, Alexis Palmer, Simon Ostermann and Manfred Pinkal. “Paraphrase Detection for Short Answer Scoring.” (2014).

[6] Magnolini, Simone. (2014). A survey on paraphrase recognition. CEUR Workshop Proceedings. 1334. 33-41.

[7] Kanjirangat, Vani & Gupta, Deepa. (2016). ASE@DPIL-FIRE2016: Hindi Paraphrase Detection using Natural Language Processing Techniques & Semantic Similarity Computations.

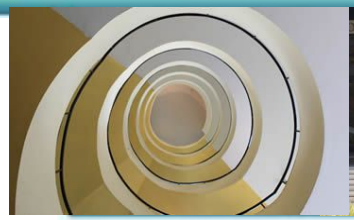
[8] Bhargava, Rupal & Baoni, Anushka & Jain, Harshit & Sharma, Yashvardhan. (2016). BITS_PILANI@DPIL-FIRE2016:Paraphrase Detection in Hindi Language using Syntactic Features of Phrase.

References

[9] P. Vigneshvaran, E. Jayabalan and A. V. Kathiravan, "An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine," 2014 International Conference on Intelligent Computing Applications, 2014, pp. 431-434, doi: 10.1109/ICICA.2014.94.

[10] A. A. Aziz, E. C. Diamal and R. Ilyas, "Paraphrase Detection Using Manhattan's Recurrent Neural Networks and Long Short-Term Memory," 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2019, pp. 432-437, doi: 10.23919/EECSI48112.2019.8976951.

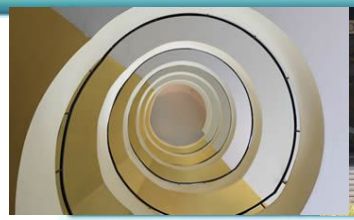
PES UNIVERSITY






Project Demo

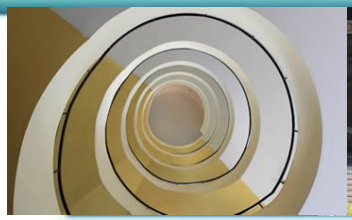
PES UNIVERSITY






Queries

PES UNIVERSITY






Thank You