Mini-Project Progress Review 2

Project Title : Paraphrase Detection in A low resourced language : Kannada

Project Guide : Mamatha H R

Project Team :

Name	SRN
Anagha H M	PES1UG19CS057
Karthik Sairam	PES1UG19CS210

Project Abstract and Scope

"Paraphrase identification in Kannada using NLP"

- Paraphrase identification is a natural language processing problem that involves the determination of whether two text segments have the same meaning.
- Our goal is to implement this in the Kannada language.

Suggestions from Review - 1

- We were told to edit the title of the project.
- Suggested that we make our own dataset.

Project Requirements

Functional requirements:

1) Ability to input texts in the Kannada language
2) Ability to output the binary classification result

Non-functional requirements:

1) To handle input in the kannada language
2) To clean the sentences, extract the features from the same.
3) For the model to scale well with increase in dataset size.

Detailed Literature Survey

Paper Title: Paraphrase plagiarism identification with character-level features

Authors: Srivastava, Shruti & Govilkar, Sharvari

This paper explains the approach of Natural Language Processing in English, and provides information on different Paraphrase detection techniques and Similarity Metrics used in NLP. In addition, the different features to be extracted are also detailed. 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.

Paper Title: Detection of paraphrases for Devanagari languages using support vector machine

Authors: D. S. Bhole and S. S. Patil

- This paper details about the feature extraction portion of paraphrase detection, including Tokenization, Stop-word elimination, Stemming and Synonyms Matching.
- Done for Hindi and Marathi, and since Kannada is significantly grammatically complicated than hindi, a more robust model is required to classify paraphrases.

Technologies Used

Technologies used :

- Python 3.8
- Pandas
- Numpy
- scikit-learn (for classification)
- NLTK (Natural Language Toolkit)

Project Progress So far

- We have completed the literature survey.
- We have completed the dataset construction. (major portion)
- Preprocessing and tokenization is done.
- Introduction, motivation and literature survey recorded in IEEE format.

Conclusion

1) The dataset construction has been finished. This includes over 500 tuples of data items in the Kannada language as paraphrases.
2) The dataset thus created has been cleaned for unwanted characters like %, “, - etc.
3) Tokenization has been completed using the INLTK module, and the tokens are kept ready for further stemming and lemmatization.

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).

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.

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.

Thank You