Natural Language Processing
Project Proposal

Mohit Jain
Kapil Thakkar

1 Goal

Building system for tweet normalisation.

2 Description

Due to restriction of only 140 characters per tweet, users often use shorthand notations. But this messages can not be processed directly in NLP field, so we may need to convert it into some consistent form. This process is called as Tweet Normalisation. Aim of our project is to take Tweets as input and find output such shorthand words and normalise it. Ke et al [1] have used syllable based method for tweet normalisation and have achieved F score of 86.08 for Lexnorm 1.2. We would be using this as our baseline.

3 Dataset

- Lexnorm 1.2

- Lexnorm2015

Apart from aforementioned dataset we would be using corpus of in vocabulary words.

4 Evaluation

Standard F-score would be used for evaluation.

References

[1] Xu, K., Xia, Y. and Lee, C. (2015). Tweet Normalization with Syllables. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, July 26-31, 2015.