# Normalization of SMS Text

Gurpreet Singh 10277 Sachin Yadav 10621

{gurpreet, sachinky} @iitk.ac.in

## INTRODUCTION

hav bin reali happy since tlkin 2 u dnt no y tho

↓

Have been really happy since talking to you Dont know why though


### Types of ill-formed words:

1. Typos (e.g. luv )
2. ad hoc abbreviations (e.g. ppl, u)
3. Phonetic substitutions (e.g. 2morrow,10q), etc.

**Problem Introduction:**

- Real-time, short, massive in volume
- Noisy
- Similar to chat text as in facebook, twitter

**Why is it Important?**

1. Text to Speech Conversion
2. Language Translation

## PREVIOUS WORK

Noisy channel model

Given ill-formed text T and standard form S, find argmax P(S|T) via argmax P(T|S)P(S), where P(S) = language model and P(T|S) =error model [Toutanova and Moore, 2002, Choudhury et al., 2007,Cook and Stevenson, 2009]

Phrasal statistical machine translation

SMT: original text = source language; normalised form = target language

[Aw et al., 2006, Kaufmann and Kalita, 2010]

Miscellaneous

Automatic Speech Recognition [Kobus et al., 2008],

Hybrid models[Beaufort et al., 2010]

## OUR APPROACH

- Separate OOV(Out Of Vocabulary) words using PyEnchant library .
- Confusion Set Generation.
- Detection of ill-formed words.
- Candidate Selection

**Confusion Set Generation:**

- Used Double Metaphone Algorithm(Phonemic Matching).
- Used difflib python module(Lexical Matching).

**Detection of ill-formed words:**

- Dependency parsing of text messages.
- Classify using SVM classifier.

**Candidate Selection:**

Exploit lexical edit distance, phonemic edit distance, prefix substring, suffix substring, and the longest common subsequence (LCS) to capture morphophonemic similarity.

```
C:\Python27_1>python test.py
Enter your text: he love to tlk abt ur styl
The word he is in Dictionary

The word love is in Dictionary

The word to is in Dictionary

Confusion Set of tlk:
['talk', 'talc']


Confusion Set of abt:
['abut', 'abet', 'about', 'abbot', 'abate']


Confusion Set of ur:
['our', 'your', 'urea', 'euro', 'aura']


Confusion Set of styl:
['styli', 'style', 'stool', 'stole', 'still']
```

He love to tlk abt ur styl

↓

He love to talk about your style

## REFERENCES


- Bo Han and Timothy Baldwin

Lexical Normalization of Short Text Messages: Make Sense a #twitter

- Deana L. Pennell and Yang Liu

A Character-Level Machine Translation Approach for Normalization of SMS Abbreviations

- Toutanova and Moore, 2002, Choudhury et al., 2007,Cook and Stevenson, 2009
- Kaufmann and Kalita, 2010
- http://images.sodahead.com for picture.
- https://github.com/dracos/double-metaphone
- http://pythonhosted.org/pyenchant/
- http://docs.python.org/2/library/difflib.html
- http://ww2.cs.mu.oz.au/~hanb/emnlp.tgz
- /usr/share/dict for english dictionary words.


## ACKNOWLEDGEMENTS

We are very thankful to :

- Prof. Amitabha Mukerjee, Computer Science and Engineering Department, IIT Kanpur.
- Bo Han and Timothy Baldwin, NICTA Victoria Research Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne
- Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu.