

Normalization of SMS Text

Gurpreet Singh Khanuja

Sachin Yadav

Problem Introduction

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

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



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

e.g., earthquak,
earthquick, ...
could be normalised
to
“earthquake”.



2525magicgirl magicgirl

This paper will discuss how to rebuild the damaged area by Tohoku **earthquick** and tsunami.

12 hours ago



ehtscindy kluht\$z;

i thought there was a **earthquick** :- #smh.

28 May



BeliebesJBiebs Verified Belieber ✓

@belieber_bella lol I wanna see a tornado in real life :(And an **earthquick** or a tsunami ! It will be cool ..Except the having chances of

28 May



CatherineNotoji ✓ Verified Belieber

Called tmnet just now.They said that cause of japan **earthquick** so the international network, streamyx broke down :-

27 May



killer_ember Marcell

-spam some nudges- OMG **EARTHQUICK!!!** LMFAO
LOLLLLLLLLLL

Objective

- To convert ill-formed English words to their standard forms.
- Typos e.g. luv
- ad hoc abbreviations e.g. ppl, u
- Phonetic substitutions e.g. 2morrow,10q etc.

- The normalization focuses on Out-Of-Vocabulary (OOV) words, which are firstly identified.

He love to **tlk abt ur styl**



He love to **talk about your style**

- Noisy channel model

Given ill-formed text T and standard form S , find $\text{argmax } P(S|T)$ via $\text{argmax } P(T|S)P(S)$, 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]

Proposed Approach

- Confusion set generation (finding correct candidates)
... Raju reach home **b4** 12 midnite ...



Confusion
set
generation



before
four
be
bore

- Detection of ill-formed word by using dependency parser(is the candidate an ill-formed word?)

before

Raju reach home ? 12 midnite

four

bore



Ill-formed word detector



Yes or No

Normalization of ill-formed word
(select canonical lexical form)

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

- <https://twitter.com>