Normalisation of SMS Text

Gurpreet Singh Khanuja, Sachin Yadav
{gurpreet,sachinky}@iitk.ac.in
Advisor : Dr. Amitabh Mukerjee
Department of Computer Science and Engineering,
IIT Kanpur, India

April 18, 2013

1 Motivation

As we know that mobile phone as most widely used now a days and are most commom mode of communication and SMS(Short Message Services) is an important part of communication through mobile phones. Its the cheapest mode of communication and is emerging as a most frequent used channel of communication among youngsters in there preferable language and slang. So we proposed to work on a project based on SMS information extraction using NLP(Natural Language Processing) and other tools in the field of shopping. Also in this modern era, people don't want to go for shopping of grocery items in which do not depend on their individual choices. People always expect a good home delivery system, but conveying the things to the shopkeeper is an overhead. As most the people have acces to mobile phones and SMS and rather than internet, so an SMS based shopping system is a very solution to their problems. Thus people can go for purchasing their products which are independent of their choices and that could be easily available to them via home delivery through SMS shopping(which would be on their fingertips). There already been work done on SMS based quering system which focus on the structure and functions of a system for querying information and knowledge by the use of SMS text messages, which can convey small pieces of materials related to learning processes, such as items of glossary, small pieces of short course summaries or examination preparation notes, student guidance, answers to exercises, second language learning tips etc. in a mobile learning environment. [2]In recent years, research in natural language processing has increasingly focused on normalizing SMS messages. [3]Actually today's written text in SMS is a major deviation from the norm of a language. To effectively process these messages, it is thus necessary to develop robust language processing tools, capable of bearing with the extreme form of noise they contain. Thus normalizing the SMS messgaes is important.

2 Our Implementation

Our implementation involves the following steps:

- Separation of OOV(Out of Vocabulary) Words.
- Confusion Set Generation.
- Detection of ill-formed words.

- Candidate Selection.

2.1 For example

hav bin reali happy since tlkin 2 u dnt no y tho
is converted to
have been really happy since talking to you dont know why though

2.2 Separation of OOV(Out of Vocabulary) Words

OOV(Out of Vocabulary) words are those words which are not present in the dictionary. For our implementation, we have used the **PyEnchant** library of python which use the standard 'en_US' dictionary for separating the OOV words from IV(In Vocabulary) words.

2.3 Confusion Set Generation

In this step, we first create a set of correction candidates called **Confusion Set** for the given OOV word which we selected in step one on the basis of morphophonemic similarity and lexical similarity.

2.3.1 Double Metaphone Algorithm

The Double Metaphone phonetic encoding algorithm can return both the primary as well as the secondry code for a given string.

2.3.2 Levenshtein Distance

Levenshtein Distance between any two strings is defined as the minimum number of edits which we can do to transform one string into another and all the edit operations that can be done are insertion, deletion and updation.

2.4 References

- Bo Han and Timothy Baldwin 2011
  Lexical Normalization of SMS text :Makn sense a #twitter

- Lawrence Philips. 2000. The double metaphone search algorithm. C/C++ Users Journal, 18:3843.

- AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normal- ization. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguis- tics, pages 3340, Sydney, Australia.

- Joseph Kaufmann and Jugal Kalita. 2010. Syntactic nor- malization of Twitter messages. In International Con- ference on Natural Language Processing, Kharagpur, India