Noisy Text Analytics

L. Venkata Subramaniam, IBM Research India

Text produced in informal settings (email, blogs, tweet, SMS, chat) and text which results from processing (speech recognition, OCR, machine translation, historical text) is inherently noisy. This tutorial will cover the efforts of the computational linguistics community in moving beyond traditional techniques to contend with the noise.

1. Overview

Text produced by processing signals intended for human use is often noisy for automated computer processing. Digital text produced in informal settings such as online chat, SMS, emails, tweets, message boards, newsgroups, blogs, wikis and web pages contain considerable noise. Also processing techniques like Automatic Speech Recognition, Optical Character Recognition and Machine Translation introduce processing noise. People are adept when it comes to pattern recognition tasks involving typeset or handwritten documents or recorded speech, machines less-so.

Noise can manifest itself at the earliest stages of processing in the form of degraded inputs that our systems must be prepared to handle. Many downstream applications use techniques meant for clean text. It is only recently that with the increase in noisy text, these techniques are being adapted to handle noisy text. This tutorial will focus on the problems encountered in analyzing such noisy text coming from various sources. Noise introduces challenges that need special handling, either through new methods or improved versions of existing ones. For example, missing punctuation and the use of non-standard words can often hinder standard natural language processing techniques such as part-of-speech tagging and parsing. Further downstream applications such as Information Retrieval, Information Extraction and Text mining have to explicitly handle noise in order to return useful results. Often, depending on the application, the noise can be modeled and it may be possible to develop specific strategies to immunize the system from the effects of noise and improve performance. This tutorial will cover:

* Various sources of noise and their characteristics as well as typical metrics used to measure noise.
* Methods to handle noise by moving beyond traditional natural language processing techniques.
* Methods to overcome noise in specific applications like IR, IE, QA, MT, etc.

2. Outline

The tutorial will have three parts:

* What is Noise
    o Detecting Noise

o Classifying Noise
o Quantifying Noise

* Processing and/or Correcting Noise
o Spelling Correction
o Natural Language Processing of Noisy Text: Segmentation, Parsing, POS
o Learning underlying language models in presence of noise

* Effect of Noise on Downstream Applications
o Information Retrieval from Noisy Text
o Information Extraction from Noisy Text
o Classification of Noisy Text
o Summarization of Noisy Text
o Machine Translation of Noisy Text

3. Target Audience

This tutorial is designed for students and researchers in Computer Science and Computational Linguistics. Elementary knowledge of text processing is assumed. This topic is expected to be of wide interest given its relevance to the computational linguistics community. Since noisy data is also the main theme of NAACL HLT 2010, good audience participation can be expected.

4. Speaker's Bio

L Venkata Subramaniam manages the information processing and analytics group at IBM Research – India. He received his PhD from IIT Delhi in 1999. His research focuses on unstructured information management, statistical natural language processing, noisy text analytics, text and data mining, information theory, speech and image processing. He often teaches and guides student thesis at IIT Delhi on these topics. He co founded the AND (Analytics for Noisy Unstructured Text Data) workshop series and also co-chaired the first three workshops, 2007-2009. He was guest co-editor of two special issues on Noisy Text Analytics in the International Journal of Document Analysis and Recognition in 2007 and 2009.