

Classification of Crowdsourced Text Correction

Megha Gupta, Dr. Haimonti Dutta (Advisor)

Second Year Annual Presentation, 2013-2014
IIIT Delhi

18-September-2015

Outline

- 1 Courses and TA duties
- 2 Conferences and Workshops
- 3 Motivation
- 4 Problem
- 5 Architecture of the Proposed System
- 6 Methodology
- 7 Results
- 8 Conclusion
- 9 Future Work
- 10 Ongoing Work

Courses and TA duties

- Courses (SGPA: 6, CGPA: 7.4)
MTH 505 - Linear Optimization (4 credits)
- TA duties
CSE 561 - Probabilistic Graphical Modelling by Chetan Arora

Conferences

- Presented in ACM India SIGKDD Conference on Data Sciences, Bangalore (IKDD CODS, 2015)
- Attended the 3rd International Conference on Big Data Analytics at IIT Delhi (BDA, 2014)

- More than 200 million paper books are being published every year.

Motivation

- More than 200 million paper books are being published every year.
- Ebooks require less storage, shared online, digitally processed, searched, translated, edited and annotated.

Motivation

- More than 200 million paper books are being published every year.
- Ebooks require less storage, shared online, digitally processed, searched, translated, edited and annotated.
- OCR results depend on factors like input paper quality, column layout, font sizes and style.

Note

OCR is the process of transforming typewritten text into machine-readable text and it is far from perfect.

Motivation

- More than 200 million paper books are being published every year.
- Ebooks require less storage, shared online, digitally processed, searched, translated, edited and annotated.
- OCR results depend on factors like input paper quality, column layout, font sizes and style.
- System answers to questions like, “What different kinds of corrections are done by users ?” or “What are the most common mistakes made by the OCR device ?”.

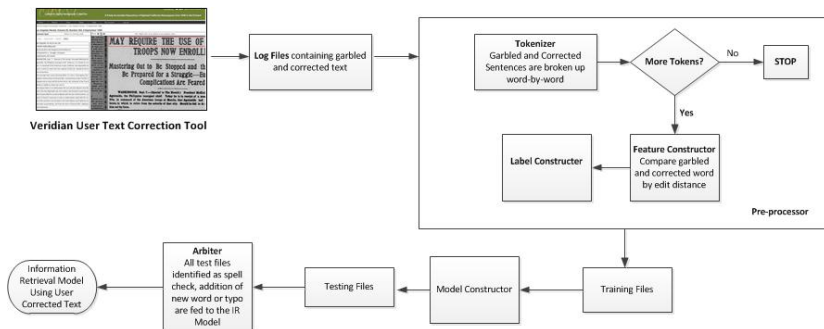
Note

OCR is the process of transforming typewritten text into machine-readable text and it is far from perfect.

Problem

To build a system for Classification of Crowdsourced Text Correction which takes input as log files containing garbled and manually corrected OCR text, parses and tokenizes them and builds models to categorize the corrections using state-of-the-art machine learning algorithms.

Architecture of the Proposed System



Text Correction Tool

Amador Ledger, Volume 1900, Number 26, 26 January 1900

Return to viewing mode

Correct text

How to correct text

I will give a Bhort history of the

Keystone mine at Amador City, Ama-

dor county. Thus property was origin-

ally located under the names of the

Keystone, Spring Hill, Middle Spring

and , Geneva veins, supposed at the

time of their location to be separate

and distinct, but as will be seen further

on, proved to be all in one great fis-

sure. The first vein worked upon was

known as the Keystone, which rests

upon the footwall slates known as "the

slates in Dry Creek," named from the

fact that it was almost dry, except dur-

ing the rainy season, the stream run-

ning parallel with the lode, which here

was nearly north and south, in its

course. I

sterned, superintendent of the
mine, and family, expect to
latter part of next week for
week, where they will visit
and in March next they will
Alaska, where Mr. Littlefield
s charge of valuable mining
s. During their residence in
r. and Mrs. Littlefield made
ands who regret their depart-

Tell Your Sister

id complexion is an impossibility
d pure blood. the sort that only ex-
ection with good digestion, a healthy
jwels. Karl's Clover Root Tea acts
the bowels, liver and kidneys, keep-
perfect health. Price 35c and 50c.
A. Goldner, Druggist.

Twelfth Century Manners.

ing to twelfth century man-
as considered more than rude
one, man or woman, to look
into space or gaze at the sky
length of time, and staring
the street or in church was
d especially impolite. People
red to be considered well
up were enjoined not to wave
d apart in the air and not to

I will give a short history of the
Keystone mine at Amador City, Ama-
dor county. This property was origin-
ally located under the names of the
Keystone, Spring Hill, Middle Spring
and Geneva veins, supposed at the
time of their location to be separate
and distinct, but as will be seen further
on, proved to be all in one great fis-
sure. The first vein worked upon was
known as the Keystone, which rests
upon the footwall slates known as "the
slates in Dry Creek," named from the
fact that it was almost dry, except dur-
ing the rainy season, the stream run-
ning parallel with the lode, which here
was nearly north and south, in its
course.

Regarding this footwall I will de-
scribe something that is interesting to
the miner. There was found a vein
three and one half feet in width, a sec-
ond vein of five feet in width, each one
distinct in character.

I am describing here a discovery
made in the Garfield shaft at the north
end of the claim. [It might also be
interesting to state that Lucius A.
Garfield, was a member of the Senate of
California in the early days and subse-
quently elected as a Delegate to Con-

twenty-one feet and 289 feet
The silicious slates in this m
edged with fine sulphurets,
which were rich in gold.

About 1000 feet south of th
said shaft the fissure had a
584 feet. The quartz body fol
most identically with the des
given of the Garfield shaft. Th
no trace of earthquake action
this fissure. The hanging-w
proven, however, to be a vast
the slates were crushed ap
from end action, as they are
and contorted and crushed at ti
fine dust.

The crystallization was
throughout and the different
showing that the fissure mu
been opened at several different
and what was supposed to b
stone was, in fact, proved to
gentic-olovine.

I will now turn my attentio
bodies on the hanging-wall, kn
worked for a long time as the
Hill. Upon this wall there is
vein dipping at an average o
grees, with a clay-slate at its
ways irregular in form and
highly pyritic, when the lode

^a<http://veridiansoftware.com/crowdsourcing/>

Datasets ¹

- Raw OCR text (Input)
- Logfiles (Input)
- Corrected OCR Text (Obtained)

¹<https://github.com/megha89/>

- Tokenizer

- Tokenizer
- Feature Constructor
 - ① Word-level Feature Construction (Proposed)
 - ② Character-level Feature Construction (Baseline)

- Tokenizer
- Feature Constructor
 - ① Word-level Feature Construction (Proposed)
 - ② Character-level Feature Construction (Baseline)
- Label Constructor
 - ① Addition
 - ② Deletion
 - ③ Punctuation
 - ④ Capitalization
 - ⑤ Spellcheck

- Tokenizer
- Feature Constructor
 - ① Word-level Feature Construction (Proposed)
 - ② Character-level Feature Construction (Baseline)
- Label Constructor
 - ① Addition
 - ② Deletion
 - ③ Punctuation
 - ④ Capitalization
 - ⑤ Spellcheck
- Model Construction : Joachim's Multiclass SVM

- Tokenizer
- Feature Constructor
 - ① Word-level Feature Construction (Proposed)
 - ② Character-level Feature Construction (Baseline)
- Label Constructor
 - ① Addition
 - ② Deletion
 - ③ Punctuation
 - ④ Capitalization
 - ⑤ Spellcheck
- Model Construction : Joachim's Multiclass SVM
- Information Retrieval Techniques

Results

Table 1 and Table 2 show the Average Loss Error and Average Time taken by the baseline and proposed method using linear and non linear kernels respectively.

Table 1: Results using linear kernel

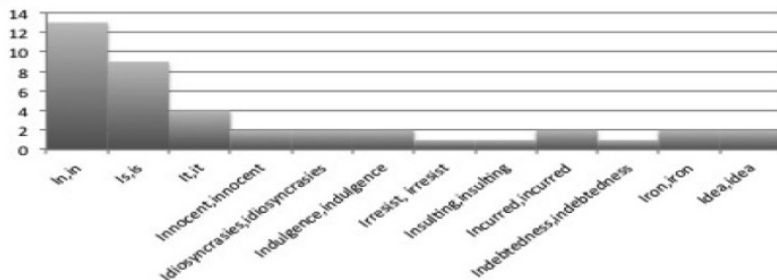
| C | AE_b | AE_p | AT_b | AT_p |
|-------|-----------|------------|------------|------------|
| .0001 | 99.43±.09 | 49.47±0.25 | 1.268±.06 | 0.11±.01 |
| .1 | 99.43±.09 | 49.464±.47 | 2.512±0.01 | 0.061±0.01 |
| 10 | 99.43±.09 | 4.974±.5 | 2.512±0.01 | 0.17 |
| 1000 | 99.43±.09 | 1.743±.14 | 3.635±0.08 | 0.382±0.04 |
| 10000 | 99.43±.09 | 0 | 6.126±0.02 | 0.303±0.02 |

Table 2: Results using polynomial and rbf kernels

| C | Polynomial | | RBF | |
|-------|------------|----------|------------|---------|
| | AE_b | AE_p | AE_b | AE_p |
| .0001 | 42±.13 | 50±.47 | 63±.15 | 27±.5 |
| 100 | 42±.13 | .33±.2 | 63±.16 | 3±.4 |
| 1000 | 42±.13 | 0±0 | 63±.16 | 1.7±.2 |
| C | AT_b | AT_p | AT_b | AT_p |
| .0001 | 37±1.4 | 10±0.2 | 25±0.7 | 6±0.2 |
| 100 | 326±11.8 | 1239±118 | 540 ± 110 | 404±34 |
| 1000 | 942±17.4 | 764±93 | 537±111.72 | 957±184 |

Results contd.

Examples of Capitalization Error, I \rightarrow i



Conclusion

- The manually crafted word-level features outperforms the automatically generated char-level features in terms of average loss error.
- The non linear kernels performed better but the time taken by them was marginally high.

Future Work

Though the manually crafted word-level dataset performed better but the time consumed was considerably higher, so there is a need to balance the tradeoff such that the algorithm becomes more compatible with the large scale datasets.

Ongoing Work

Problem: To perform aggregate load forecasting for disparate energy data sources using the ensemble based learning technique called Product of HMMs.

- Motivation

²<http://redd.csail.mit.edu/>

³<http://open.enernoc.com/data/>

Ongoing Work

Problem: To perform aggregate load forecasting for disparate energy data sources using the ensemble based learning technique called Product of HMMs.

- Motivation

- ① To avoid the unnecessary redundant information thus reducing the network traffic and improving the privacy of the customers.

²<http://redd.csail.mit.edu/>

³<http://open.enernoc.com/data/>

Problem: To perform aggregate load forecasting for disparate energy data sources using the ensemble based learning technique called Product of HMMs.

- Motivation

- ① To avoid the unnecessary redundant information thus reducing the network traffic and improving the privacy of the customers.
- ② Efficient power system planning and operation, energy purchasing and generation, load switching and infrastructure development.

²<http://redd.csail.mit.edu/>

³<http://open.enernoc.com/data/>

Problem: To perform aggregate load forecasting for disparate energy data sources using the ensemble based learning technique called Product of HMMs.

- Motivation

- ① To avoid the unnecessary redundant information thus reducing the network traffic and improving the privacy of the customers.
- ② Efficient power system planning and operation, energy purchasing and generation, load switching and infrastructure development.
- ③ Various factors that effect load forecasting are time factors, weather conditions, class of customers, special events, electricity price, fluctuating demand and supply.

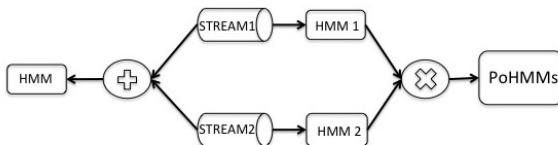
- Dataset: REDD², IIITD Faculty Housing, Enernoc dataset ³.

²<http://redd.csail.mit.edu/>

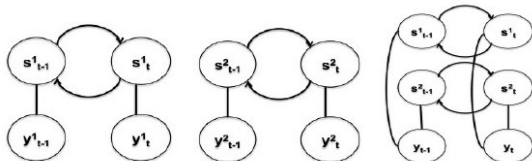
³<http://open.enernoc.com/data/>

Approach

- Load forecasting at utility level is done in 3 ways:
 - 1 completely aggregated
 - 2 completely disaggregated
 - 3 clustering based approach
- PoHMMs is a model that combines several HMMs by multiplying their individual distributions together and then renormalizing them.

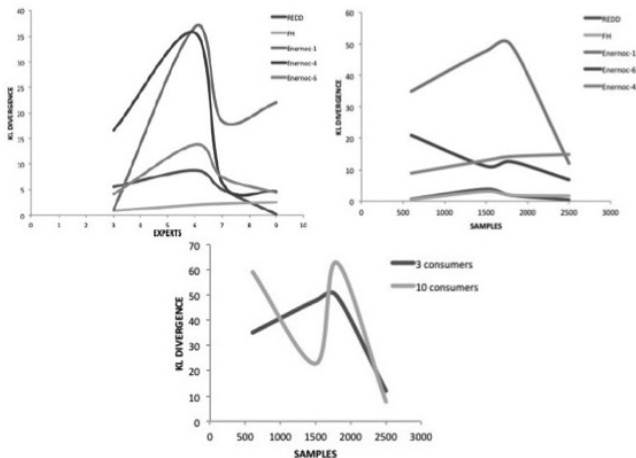


- Figure below shows two HMMs S^1 and S^2 generated by two different data streams, the aggregate energy consumption can be modelled using PoHMMs as shown below.



Results

Performance comparison between REDD, FH and Enernoc datasets is shown below.



TREC CDS Task

Problem: To retrieve the full biomedical articles that are relevant for answering generic clinical questions (“test”, “diagnosis”, “treatment”) about medical records.

Thanks
Questions?