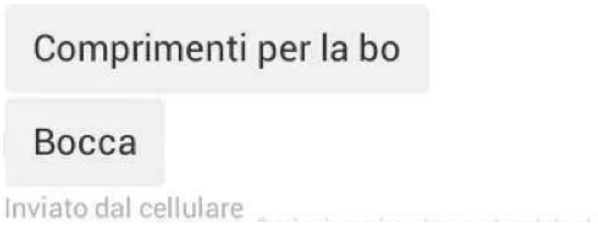
Progetto #4: MISPELLING

Motivazione

All'interno degli smartphone di nuova generazione sono disponibili strumenti automatici in grado di correggere i testi scritti all'interno dei messaggi testuali che ogni giorno vengono inviati (mediante Whatsapp) e/o diffusi sui social network (tramite Facebook e Twitter). E' stata realizzata una nuova tastiera sperimentale che sfrutta le potenzialità degli Hidden Markov Models per correggere errori e rendere digitazione e predizione delle parole più precisa che mai.

Il modello di correzione automatica: Hidden Markov Model

In questo problema, lo stato si riferisce alla lettera corretta che l'utente avrebbe dovuto digitare, e l'osservazione si riferisce alla lettera che è stata digitata sulla tastiera. Data una sequenza di osservazioni (cioè, lettere effettivamente digitate), il problema consiste nel ricostruire la sequenza di stati nascosti (cioè la sequenza prevista di lettere) che verosimilmente corrispondono alle lettere corrette.



Per definire il modello, si avranno a disposizione:

- una serie di dati di training testuali costituiti da una osservazione e dalla corrispondente una Ground Truth

OSSERVAZIONE: @martaR oggi sono ndata a vedere il nuovo negozi di Prada!!!

GROUND TRUTH: @martaR oggi sono andata a vedere il nuovo negozio di Prada!!!

Per la fase di inferenza, si avranno a disposizione solo le osservazioni (non utilizzate durante la fase di training.

Obiettivi del progetto:

1. Creazione del dataset di training e testing
   a) Raccogliere un insieme di tweet utilizzando le API di Twitter.
   b) Perturbare i testi introducendo il 10% di errore di scrittura.
   c) Dividere i testi in training (80%) e testing (20%).

2. Modello HMM

a) Definizione della struttura HMM per inferire il testo corretto, date le osservazioni derivanti dalla tastiera.
b) Stima dei parametri.

3. Correzione su nuovi testi
   a) Inferire il testo corretto, dato il messaggio inserito dell'utente.

4. Analisi dei dati
   a) Stimare le capacità predittive del modello rispetto alla ground truth.
   b) Valutare le capacità predittive del modello al crescere dei dati di training.

Software

Software utilizzabili:

- Twitter Libraries
  - https://dev.twitter.com/overview/api/twitter-libraries

- Hidden Markov Models:

  1. Python:
     a. http://ghmm.org/
  2. Matlab:
     a. http://it.mathworks.com/help/stats/hidden-markov-models-hmm.html
     b. https://github.com/probml/pmtk3
  3. Java:
     a. http://mallet.cs.umass.edu/index.php
  4. R: https://cran.r-project.org/web/packages/HiddenMarkov/index.html