CS565: Intelligent Systems and Interfaces

Lecture: Sequence Labeling or Tagging Problems

10th Feb, 2016

Semester: Jan – May 2016

Ashish Anand
IIT Guwahati

Recap

- Language Modeling
  - Parameter Estimation: Smoothing Techniques
  - Evaluation: Perplexity

- Reading Assignment
  - N-Grams [Speech and Language Processing by Jurafsky and Martin]

Moving Forward

- Sequence Labeling Problems
- Generative Models
- Hidden Markov Models

Sequence Labeling Problems

Part of Speech (PoS) Tagging

- Input: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

- Output: Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

- N: Noun; V: Verb; P: Preposition; Adv: Adverb; Adj: Adjective ….

Named Entity Recognition

- Input: "The mood in the market is very much 'sell today, ask questions later' which is a boost for Treasuries and that flight to safety is led by fear," said Gennadiy Goldberg, interest rate strategist at TD Securities in New York. [source: economictimes.indiatimes.com]

- Output: "The mood in the market is very much 'sell today, ask questions later' which is a boost for Treasuries and that flight to safety is led by fear," said [person Gennadiy Goldberg], interest rate strategist at [company TD Securities] in [location New York].

NER as Sequence Labeling Problem

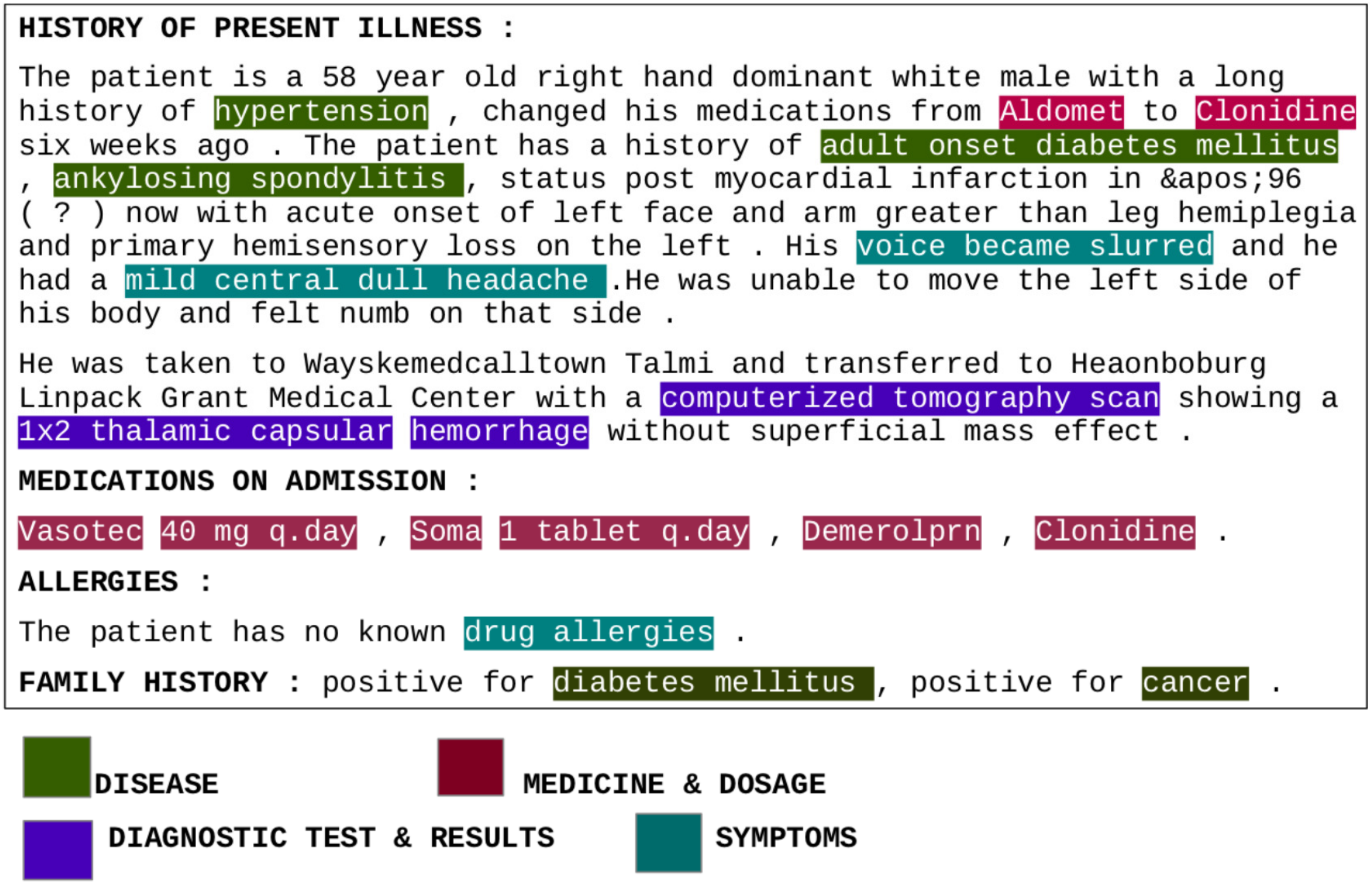
- Input: "The mood in the market is very much 'sell today, ask questions later' which is a boost for Treasuries and that flight to safety is led by fear," said Gennadiy Goldberg, interest rate strategist at TD Securities in New York. [source: economictimes.indiatimes.com]

- Output: "The/O mood/O in/O the/O market/O is/O very/O much/O '/Osell/O today/O, ask/O questions/O later/O'/O which/O is/O a/O boost/O for/O Treasuries/O and/O that/O flight/O to/O safety/O is/O led/O by/O fear/O, /O"/O said/O Gennadiy/BP Goldberg/IP, interest/O rate/O strategist/O at/O TD/BC Securities/IP in/O New/BL York/IP. /O

- O: Other; P: Person; L: Location; C: Company; B: Beginning; I: Intermediate

[BIO model]

Concept Recognition Problem



HISTORY OF PRESENT ILLNESS :

The patient is a 58 year old right hand dominant white male with a long history of hypertension , changed his medications from Aldomet to Clonidine six weeks ago . The patient has a history of adult onset diabetes mellitus , ankylosing spondylitis , status post myocardial infarction in &apos;96 (?) now with acute onset of left face and arm greater than leg hemiplegia and primary hemisensory loss on the left . His voice became slurred and he had a mild central dull headache .He was unable to move the left side of his body and felt numb on that side .

He was taken to Wayskemedcalltown Talmi and transferred to Heaonboburg Linpack Grant Medical Center with a computerized tomography scan showing a 1x2 thalamic capsular hemorrhage without superficial mass effect .

MEDICATIONS ON ADMISSION :

Vasotec 40 mg q.day , Soma 1 tablet q.day , Demerolprn , Clonidine .

ALLERGIES :

The patient has no known drug allergies .

FAMILY HISTORY : positive for diabetes mellitus , positive for cancer .

🟩 **DISEASE** 🟥 **MEDICINE & DOSAGE**

🟪 **DIAGNOSTIC TEST & RESULTS** 🟦 **SYMPTOMS**

Challenges: Ambiguities

- PoS Tagging: Words having more than one PoS
  - The back/JJ door
  - On my back/NN
  - Promised to back/VB the bill

- NER
  - Washington/BP Jr./IP

Challenges: Rare words

- Words not seen in the training data.

More formally

- Sequence Labeling or Tagging Problem
    - Definition: On board

Tagging as supervised learning problem

- Generative vs Discriminatory Models

- Generative Model
  - Hidden Markov Model (HMM)