Deecamp: NLP

Roy

VP of Data Science

AI Institute, Sinovation Ventures

创新工场 SINOVATION VENTURES

创新工场 人工智能工程院 SINOVATION VENTURES AI INSTITUTE

Agenda

- Introduction
- Case studies
  - Search engine
  - Sentiment analysis
  - Chatbot

Introduction

Task

- Understand natural signal
- Apply such understanding

Areas

- Part-of-speech (POS) tagging
- Information/entity extraction
- Sentiment analysis
- Machine translation
- Summarization
- Dialog
- Etc

Technologies

- BoW
- Regex
- Lexicons
- Word2Vec
- SVM
- Decision trees
- Deep learning
  - LSTM

Search Engine

Applications

- Query understanding
  - Query classification
  - Entity extraction and ranking
- Document understanding
  - Snippet generation
- Ranking

Music Query Classification

- Lexicons
  - Music specific: song titles, singers/bands, genres
  - FP domains: movie titles, celebrities
- Query pattern
- Highly imbalanced classes (3%)

Snippet Generation

- Summarization
  - Query word occurrences as features

All-in-One Entity Ranker

- Entity repository
  - Web crawling
  - Attribute extraction
  - Entity conflation
- Ranker training: attribute generalization
  - Title and genre for movie and music

Sentiment Analysis

Applications

- Product review
  - Amazon, Netflix, etc
- User studies prediction
  - Twitter brand ads
- Trend prediction
  - Quantitative trading

Amazon Product Review

- Challenges
  - Ambiguity
  - Coreference resolution
  - Negations
  - Sarcasm

Stock Community Comments

- Three classes
  - Positive, negative, neutral
- Feedback v.s. speculation
- Label generation
  - Manual
  - Cross-validated

Chatbot

Dimensions

- Open v.s. closed domain
- Long v.s. short conversations
- Retrieval-based v.s. generative models

Common Challenges

- Incorporating context
- Coherent personality
- Evaluation
- Intention and diversity

Current Status of Production Systems

- Requirement for high quality means:
  - Most are retrieval based
  - Most are closed domain
  - None deep-learned (though could be in the future)
- Mistakes are very costly
  - Grammar mistakes drive users away
  - Political mistakes take PR to fix

How to build a chatbot

- Define the domain (open v.s. closed)
- Get as much data as you can
- Start with retrieval-based
- Build base-line models
  - Random
  - TF-IDF
  - Sequence to sequence
  - Dual encoder LSTM
- Make an MVP and have ppl try it early

Thanks

Roy

roy@sinovationventures.com

创新工场 SINOVATION VENTURES

创新工场 人工智能工程院 SINOVATION VENTURES AI INSTITUTE