Heyday Capstone Project: Week 2 Progress

Andrew, Mrinal, Sijia, Varadraj

Table of Contents

- Project timeline
- Data preprocessing
- NER strategies considered
  - Challenges
  - Out-of-the-box NER systems
  - Rule-based systems

Project Timeline

- Week 3: Named Entity Recognition
- Week 4: Negation Detection/Scoping
- Week 5: Miscellaneous (e.g., Docker, UI, augment validation data)
- Week 6: Cleaning and formatting
- Weeks 7 & 8: Cleaning, final report, video presentations

Data Preprocessing

- Removed negative label from negative-affix words in the validation set (except when it actually was negated, e.g. “I don’t want the screw less”)

- Separately tokenized punctuation from preceding word and added O label to punctuation

- Removed “N-” from IOB labels, replacing it with a new “is_negative” column to keep NER and negation tasks separate

- Trialed a few different spelling correction packages to help downstream models; pyspellchecker most effective (especially when instructed to avoid correcting domain-specific words)

NER Task: Main Challenges

- In e-commerce domain, haven't found appropriate:
  - Gold data (i.e., annotated examples)
  - Silver data (e.g., text from Wikipedia with wiki links assumed to denote named entities)
- Training on validation set would cause leakage (and validation set is small anyway)
- Even a rule-based system designed to handle our small validation set might be "overfitted"

NER Task: Approaches

- **Out of the box NER models**
  - **Most promising: Ontonotes NER**
    - **Labels somewhat different than in validation set**
    - **On most sentences, does not detect anything**
    - **Appears able to handle comparatives well**

NER Task: Approaches

- Semi-supervised learning
  - Train on validation set, annotate plain text data, train on latter annotations, evaluate on validation set, repeat
  - Challenges:
    - Involves leakage of data from validation; but without training on the validation set, how can we annotate further examples and use them as training data?
    - The two sources of unannotated data we have considered (Amazon reviews and Wikipedia pages) each have their own problems

NER Task: Approaches

- Rule-based approaches
- We have not yet attempted to build a rule-based system, but have considered the following strategies:
  - Parsing/chunking
  - Hard-coded sentence patterns
  - Gazetteers (derived from Wikipedia, validation set, and product taxonomy provided by Haejin)
  - Word embeddings with cosine similarity