Heyday Capstone project: Week 3 Progress

Andrew, Mrinal, Sijia, Varadraj

Table of Contents

Negation detection corpus collection

- SEM 2012 Shared Task – Resolving the Scope and Focus of Negation
  - CD-SCO
    - The dataset for scope detection.
  - PB-FOC
    - The dataset for focus detection.
- SFU negation corpus

CD-SCO

- This dataset includes two stories by Conan Doyle for training and development.
- All occurrences of negation are annotated (1,056 out of 3,899 sentences), accounting for negation expressed by nouns, pronouns, verbs, adverbs, determiners, conjunctions and prepositions.
- The negation cue and scope are marked for each negation cue and the negated event if any.
- Cues and scopes may be discontinuous.

PB-FOC

- Focus of negation is annotated over the 3,993 sentences in the WSJ section of the Penn TreeBank marked with MNEG in PropBank.
- It accounts for verbal, analytical and clausal relation; the role most likely to correspond to the focus was selected as focus.
- Unlike [CD-SCO], all sentences in [PB-FOC] contain a negation.
- 80% of [PB-FOC] will be released as training/development set and the rest for test.

SFU negation corpus

- One layer with four possible annotations for each token of word: NEG, SCOPE, FOCUS, and XSCOPE.
- We would use the SCOPE annotation as our label in training, then predict negation scope on dev sentences, and tag an entity as negative if it falls within the scope
- Corpus consists of 1,043 tokens
- Example:

5-5	550-552	no	NEG
5-6	553-559	amount	SCOPE[5]
5-7	560-562	of	SCOPE[5]
5-8	563-568	money	SCOPE[5]|FOCUS[6]

Rule-based system for negation

- Still in progress; hope to have some rules by end of day (getting appropriate data structure has been complicated)
- Scope-based negation detection as in SFU corpus can't be directly applied to dev set with much success since negated entities often fall outside scope of negation
    - Partly because e.g. "instead" not counted as negator in training data
- Will apply negation functions related to each negator (e.g., no, not, etc.) to each word/entity; if the number of times an entity is negated is odd, the indices of the span of that entity will be appended to a list e.g. "Negative entities:[(1,5)]" if there is a negated entity from index 1 to 5. This prediction can be evaluated against the dev set.

Knowledge-based system

- For knowledge base creation, we collected the attributes from amazon reviews. We have about 180k attributes across 53 categories:

  ['Color:', 'Format:', 'Style Name:', 'Size:', 'Material Type:', 'Package Quantity:', 'Size Name:', 'Style:', 'Item Package Quantity:', 'Item Display Length:', 'Length:', 'Package Type:', 'Design:', 'Item Display Weight:', 'Product Packaging:', 'Number of Items:', 'Flavor:', 'Flavor Name:', 'Color Name:', 'Scent Name:', 'Display Height:', 'Pattern:', 'Shape:', 'Scent:', 'Team Name:', 'Outside Diameter:', 'Overall Length:', 'Model Name:', 'Model Number:', 'Item Weight:', 'Material:', 'Configuration:', 'Curvature:', 'style:', 'Edition:', 'Denomination:', 'Colour:', 'Offer Type:', 'Wattage:', 'Overall Width:', 'Part Number:', 'Thickness:', 'Line Weight:', 'Item Shape:', 'Gauge:', 'Hand Orientation:', 'Inside Diameter:', 'Connector Type:', 'Volume:', 'Capacity:', 'Bore Diameter:', 'Model:', 'Grit Type:']

- Potentially, this is how we are planning to use the attributes:
  - Avoiding correction of words which are part of the attributes
  - Extraction of attributes in a user request
  - Detection of product category based on the attribute type for eg. if the attribute is of type Model Name under home category, we can maybe pass this information to our NER model for improved accuracy for the product extraction
  - We also have saved the reviews into the file in case we need to use those later.

Training Data Generation using GPT2

- Generated text data using the GPT2 in order to generate domain specific training data.
- The same yielded positive outcomes.
- Some examples are shown below.

```
"Hi, do you sell fingerless gloves please"
Need an inexpensive thermostatic shower set with an extra outlet for bidet. Can you suggest a couple.
"hi, can I put a tankless toilet into a 2""x 4"" wall?"
"I am looking for a left offset vanity set up to 48 inches wide, white, brushed nickel hardware"
seamless undermount
Touchless kitchen faucet
Looking for bowl with pattern LESS than 5 inches deep. Can be 7 inches wide
Looking shower door frameless of 60 inch standing shower
Do you sell beige toilets. Bone and buscuit are too light in colour
We're looking for a shower door set greater than 70 inches wide
18 inch deep by 36 inches long brown freestanding vanity
Looking for a 36 inch vanity with granite top
I am looking for bathroom mirrors black framed in rectangular and oval
looking at 36 inch medicine cabinet for $399 . is it recessed able to fit into wall as a flush mount?
Vessel bag in goldish color
Looking for protein powder no carb no suger
"I am brand new to fat burners, I want to know a good fat burner to see good results."
i want protein without lactose
I'm looking for whey
```

BERT for token classification

- Token classification was done using BERT.
- The outcomes were not very favourable.
- With the 'O' tag included, the validation accuracy was roughly 67 percent.
- BERT is able to detect the SIZE and COLOUR attribute with minimal data and with a recall of 0.32 and 0.20.
- Tried BERT due to the unfamiliarity with flair modelling.
- Also tried passing single sentences into the flair/ner-english-ontonotes-large,but the results were not promising. Except for the product's size property, the model was unable to recognise any other tags.
- In order to fine tune to the problem statement, BERT requires more tagged training data.
- Will try using Rule-Based and CRF models for token classification

References

- *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation: https://aclanthology.org/S12-1035.pdf
- Dataset of scope and focus of negation: https://www.clips.uantwerpen.be/sem2012-st-neg/#Datasets
- Multitask Learning of Negation and Speculation using Transformers: https://aclanthology.org/2020.louhi-1.9.pdf
- What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis: https://aclanthology.org/W10-3110.pdf
- NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution: https://arxiv.org/pdf/1911.04211.pdf
- SFU Opinion and Comments Corpus (SOCC): https://www.kaggle.com/datasets/mtaboada/sfu-opinion-and-comments-corpus-socc?resource=download

THE UNIVERSITY OF BRITISH COLUMBIA

Thank you !