Assignment Discussion #3 and Project Discussion

Ashwani Kumar Jha 20305R001
Lackarsu Siri Verma 180050052
Mikhil Gupta 203079016

12th October, 2021

Assignment Discussion #3

WSD continuation, NEI

(WSD template already given; use that)

Performance report of HMM-WSD

- Precision - 4.054
- Recall - 4.037
- F1-score - 4.053

Confusion Cases (for HMM)

It urged that the city take steps to remedy this problem

steps:

- Actual tag: ‘step.n.01’ : any maneuver made as part of progress toward a goal
- Predicted tag: ‘step.n.03’ : the act of changing location by raising the foot and setting it down

Nevertheless, we feel that in the future …

feel:

- Actual tag: ‘feel.v.02’ : come to believe on the basis of emotion, intuitions, or indefinite grounds
- Predicted tag: ‘feel.v.01’ : undergo an emotional sensation or be in a particular state of mind

Only a relative handful of such reports was received, …

only:

- Actual tag: ‘only.r.01’ : and nothing more
- Predicted tag: ‘only.r.02’ : without any others being included or involved

handful:

- Actual tag: ‘handful.n.01’ : a small number of amount
- Predicted tag: ‘handful.n.02’ : the quantity that can held in the hand

received:

- Actual tag: ‘receive.v.02’ : receive a specified treatment (abstract)
- Predicted tag: ‘receive.v.01’ : get something, come into possession of

Interpretation of confusion (error analysis: HMM)

- Misclassification of Grouped Words-In most of the cases HMM model was not able to correctly predict the sentence for group word.

  Eg- NE Fulton county Grand Jury
- Here we are only considering Bigram assumption,but actual sense depends on whole sentence context.
- In some cases, the POS tag of the word is predicted incorrectly, further adding to the error.

Data Processing Info (Pre-processing: HMM)

- Calculated the index of test and train data for each fold validation.
- Lowercase the sentence
- Added ‘^’ as first word/tag in each sentence
- Calculated frequency of emission_matrix[tag][word] and transition_matrix[prev_tag][tag] based on the training data
- Calculated probability of emission_matrix[tag][word] and transition_matrix[prev_tag][tag] based on the calculated frequency
- Created a dictionary to store tags for a particular word

Performance report of Word Vector Based Overlap approach

- Precision - 0.2972
- Recall - 0.2940
- F1-score - 0.2955

Interpretation of confusion (error analysis: WE-Overlap)

1. In the sense tagged dataset, some senses have been given to a group of words instead of a single context word, and these senses are not usually part of the synsets of the words.
2. For named entities(words starting with NE), the actual tags are `['group.n.01', 'location.n.01', 'person.n.01']`, however these tags are not included in the senses of the named entities.

Interpretation of confusion (error analysis: WE-Overlap)

1. While calculating the sense embedding of senses, the word2vec embeddings of functional words in the gloss are also included. These words can sometimes give more meaning to irrelevant senses than the relevant ones.
2. About 2.6% words in the Semcor dataset have been tagged as NE instead of a specific sense.

Data Processing Info (Pre-processing: WE-Overlap)

- The whole dataset is taken as the test data.
- For each sentence in the test data =>

  context_emb = sum of word vector of each word in input /# of words in input
- For each word in the sentence, there are a list of senses associated with it, for which sense embedding vectors are calculated.

  sense_emb = sum of word vector of each word in Gloss /# of words in Gloss
- Using the cosine similarity between sense embedding and context embedding for each sense of a word, the best sense is picked for a word in a given input sentence.
- This is repeated for all the sentences in the test data, and the predicted sense tags are recorded.
- To compare the predicted and actual tags, the detailed lemma tag of words in the dataset have been converted to synset tags.

Problem Definition: NEI

- Given a sentence/document, mark each token as 1/0 as per whether the token is a Named Entity or not
- If the named entity consists of multiple words just continue with 1s until a non-NE appears
- E.g. *The_0 State_1 Bank_1 of_1 India_1 is_0 the_0 largest_0 bank_0 in_0 the_0 country_0 . _0*

DATA

- CoNLL 2003
- train : test : valid =219554:50350:55044
- train sentences = 14041
- Named Entity = 34043
- Non Named Entity = 169578

Feature Engineering

- Features: Length of the word , POS, Prev-POS, Next-POS ,Initial Capital letter (bool) , All cap letter (bool)
- we created a pos to num mapping to store the tag value for POS/Prev-POS and Next POS
  - we used SVM classifier.
  - Kernel : rbf , poly , sigmoid , linear

Justification of Feature Set

- POS play important role in NEI classification
- The Structuring of sentence i.e prev and next tag to decide nei (dependence on)
- Structure of the word (initial or whole caps) can be used to classify the name entity word

Performance

- P, R, F, Accuracy
- precision :0.8912799706500112
- recall :0.8923007281100408
- f1score: 0.891790057286208
- f2score:0.8920963896137768

Confusion Matrix

Predicted □ Actual (rows)	0	1
0	41225	1562
1	1534	7041

Result Interpretation

- The Non NEI classified as the NEI for some words due to the first cap letter at the starting of the sentence
- Dataset is linear
- The order of importance of the feature is :

  POS>prev_pos>next_pos>all_caps >first_cap > length

Project

Sentiment Analysis for Code-Switched Languages

Problem Statement

- Given tweets in Code Switched Language, Output the sentiment expressed. Sentiment can be positive, negative or neutral

- Input-“This restaurant is nice. Khana lazawab hai yaha.”
  output- Positive(1)

- Data collected from tweets by using the list of tokens of hindi words released by Patra et al[2018].

Why is the problem important

- Sentiment Analysis is a hot topic of research due to availability of opinions of users on social media and other platforms.

- Users post their opinion as per their ease -
    a. “@prahladspatel modi mantrimandal me shamil hone pr hardik badhai”
    b. “This restaurant is nice. Khana lazawab hai yaha.”
- In the second example, a user posted his opinion by using a code-mixed language, in this case Hindi and English together, popularly known as HingLish

- we will analyse the sentiment when the feedback is given in code switched mode (English and Hindi)

What is hard about the problem

- The main challenge about the problem is the switching of languages between the sentence.
- Sarcasm recognition
- misspelled hindi words in the sentence/Incomplete sentences

What has been done on this problem so far

Dataset statics-Data is collected from tweets by using the list of tokens of hindi words released by Patra et al[2018].

split	Total	positive	neutral	negative
Train	14000	4634	5264	4102
Test	3000	982	1128	890
validation	3000	1000	1100	900
Total	20000	6616	7492	5892

Dataset Pre-processing-

Data-set -> Total 14k sentences (evenly distributed positive,negative and neutral)

Performed preliminary data processing operations like removal of unwanted links,special characters,etc and basic cleaning, expansion of short forms,stemming etc.

Preprocessing

- Used the English/ Hindi tagged words to segregate chunks of hindi words from chunks of english words
- use of Google translator for converting the hindi slangs and chunk of words with english data for better processing.
- Sentiment Intensity Analyser: to calculate the intensity of negative , positive and neutrality in the sentence
- we will Process hindi slang words separately using Offensive hindi tweet dataset
- After pre-processing the data and before feeding into the actual model,data will have following feature-

  1. code mixed sentence 2.review translated into english language 3.whether slang exists in data or not? 4.Negative score of the sentence 5.Positive score of the sentence or review 6.neutral score of the sentence 7.offensive rating of the slang

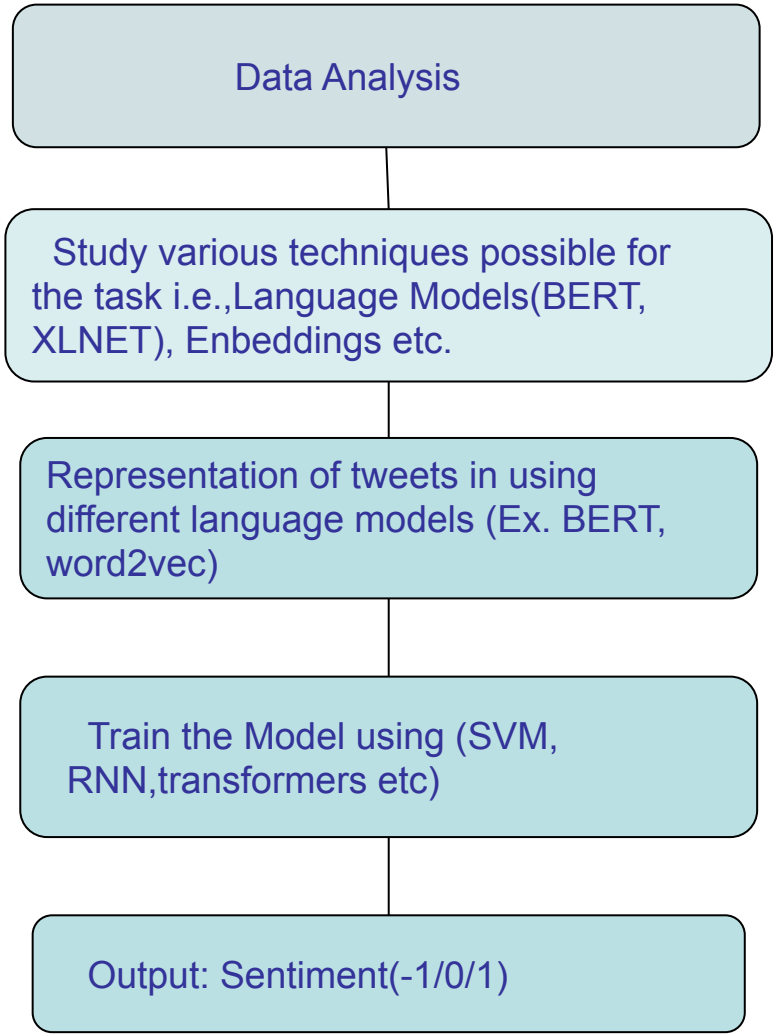
Performance Analysis-

- Report the accuracy for language Identification (LId) in the given dataset
- Compare accuracies obtained using different models (SVM, RNN, Transformers) using different language models for the task of sentiment analysis

- Detailed Error Analysis for the work done

Your tackling of the problem

Methodology



References

- Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. **Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task @icon-2017.** CoRR, abs/1803.06745.

- Patwa, Parth and Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL,Srinivas and Gamb'ack, Bj'orn and Chakraborty, Tanmoy and Solorio, Thamar and Das, Amitava, **SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets**, Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), December , ACL 2020

Thank You