Project Tips and Guidance

Natural Language Processing

Suggested Readings

1. Practical Methodology (Deep Learning book chapter)

Project Deliverables

21 Students > 7 Group of 3 members:

Final Project (45%)

1. Novelty (related work) (20%)
2. Experiment rigour (comparisons) (20%)
3. Model complexity (competency) (20%)
4. Evaluation methods (appropriate) (20%)
5. Effort (not last day!) (20%)

Submission deliverables:

1. Python file (e.g., notebook, .py)
2. Presentation file (e.g., .pdf, .ppt)

Timeline

8 March : Today
10 March : First Group-Meeting
15 March : Recent NLP Trend
17 March : Q&A
22 March : Progress Day
24 March : Progress Day
29 March : Q&A
31 March : Q&A

5 April : Q&A
7 April : Final Project Presentation
12 April : Final Project Presentation
19 April : L18 Knowledge Integration
21 April : L19 Coreference Resolution

Practical Tips

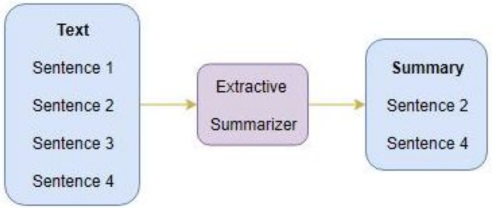
Practical Tips

- Pretrained models can help a lot (Huggingface)
- Read a lot of papers

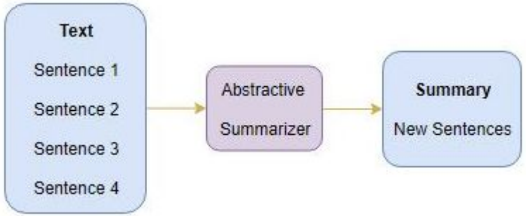
Text Summarization

Text Summarization

Extractive



Abstractive



Problems existing in abstractive methods

1. Model Performance
2. Factual Inconsistency
3. Understanding of Models

Model Performance

- How to improve distant dependency?
- Self attention is computationally expensive O(n2)
- How to extend number of input token?
- Exploring different types of masking

Factual Consistency Evaluation Metric

- How to eliminate Name Entity Hallucinations?
- Is n-gram evaluation enough? Misinterpretation?

Understanding of Models

- Why model performs particularly well on certain samples?
- Why model performs particularly poorly on certain samples?
- Explore features those samples

Datasets

Xsum

- BBC articles (2010 to 2017)
- covers a wide variety of domains (e.g., News, Politics, Sports, Weather, Business, Technology, Science, Health, Family, Education, Entertainment and Arts)
- Train: 204,045 (90%)
- Validation: 11,332 (5%)
- Test: 11,334 (5%)

CNN/DailyMail

- news stories in CNN and Daily Mail websites
- Train: 286,817
- Validation: 13,368
- Test: 11,487

Papers

Model Performance

- Attention is all you need https://arxiv.org/abs/1706.03762
- Longformer https://arxiv.org/abs/2004.05150
- BigBird https://arxiv.org/abs/2007.14062
- T5 https://arxiv.org/abs/1910.10683
- Pegasus https://arxiv.org/abs/1912.08777

Factual Consistency

- BERTScore https://arxiv.org/abs/1904.09675
- FactCC https://arxiv.org/abs/1910.12840
- Entity-level https://arxiv.org/abs/2102.09130
- SUMMVIS https://arxiv.org/abs/2104.07605

Understanding of Model

- Pathology https://arxiv.org/abs/1804.07781
- What does BERT look at https://arxiv.org/abs/1906.04341
- Knowledge T5 https://arxiv.org/abs/2002.08910

Intent Detection Using Pre-Trained Models

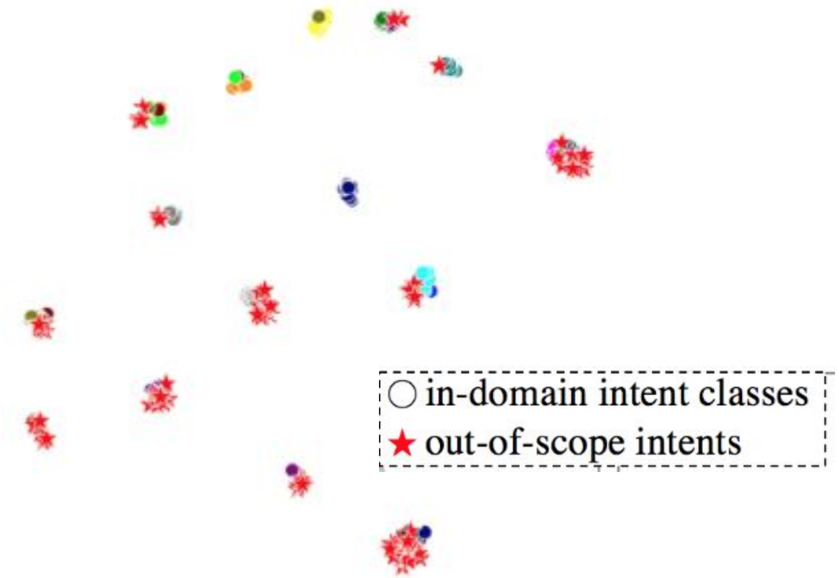
Outline

- Why this is important ?
- Problem existing these datasets
- Model performances
- Related works

Why intent detection task is important ?

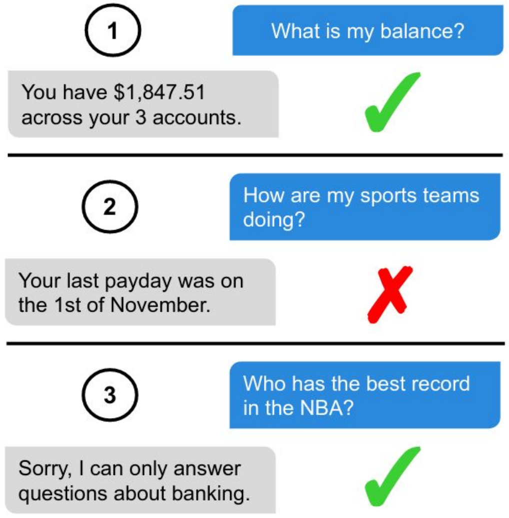
What is intent detection ?

Problems



tSNE of embeddings before their classifier layers are used

https://arxiv.org/pdf/2010.13009.pdf



Example exchanges between a user (blue, right side) and a task-driven dialog system for personal finance (grey, left side)

https://arxiv.org/pdf/1909.02027.pdf

Pre-training Datasets

Name	# Utterance	# Intent	# Domain
CLINC150 (Larson et al., 2019)	18200	150	10
BANKING77 (Casanueva et al., 2020)	10162	77	1
HWU64 (Liu et al., 2019)	10030	64	21
TOP (Gupta et al., 2018)	35741	25	2
SNIPS (Coucke et al., 2018)	9888	5	-
ATIS (Tur et al., 2010)	4978	21	-

Table 1: Data statistics for intent detection datasets.

Evaluation Datasets

Model	CLINC150 5-shot	CLINC150 10-shot	BANKING77 5-shot	BANKING77 10-shot	HWU64 5-shot	HWU64 10-shot
RoBERTa+Classifier (Zhang et al., 2020a)	87.99	91.55	74.04	84.27	75.56	82.90
USE (Casanueva et al., 2020)	87.82	90.85	76.29	84.23	77.79	83.75
CONVERT (Casanueva et al., 2020)	89.22	92.62	75.32	83.32	76.95	82.65
USE+CONVERT (Casanueva et al., 2020)	90.49	93.26	77.75	85.19	80.01	85.83
CONVBERT (Mehri et al., 2020a)	-	92.10	-	83.63	-	83.77
CONVBERT + MLM (Mehri et al., 2020a)	-	92.75	-	83.99	-	84.52
CONVBERT + Combined (Mehri et al., 2020b)	-	93.97	-	85.95	-	86.28
DNNC (Zhang et al., 2020a)	91.02	93.76	80.40	86.71	80.46	84.72
CPFT	**92.34**	**94.18**	**80.86**	**87.20**	**82.03**	**87.13**

Table 2: Testing accuracy (×100%) on three datasets under 5-shot and 10-shot settings.

Related Works

Intent Detection

Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning : https://arxiv.org/pdf/2109.06349.pdf

Discriminative Nearest Neighbor Few-Shot Intent Detection by Transferring Natural Language Inference: https://ieeexplore.ieee.org/document/8998086

Efficient Intent Detection with Dual Sentence Encoders : https://arxiv.org/abs/2003.04807

Explainability

What Does BERT Look At? An Analysis of BERT's Attention: https://arxiv.org/abs/1906.04341

Emergent linguistic structure in Ann trained by self-supervision : https://arxiv.org/abs/1909.02027

Visualizing and understanding Recurrent Networks: https://arxiv.org/abs/1506.0207

Dataset Evaluation

An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction: https://arxiv.org/pdf/1909.02027.pdf

Depression Detection In Social Media

Dataset

	Depression	Control
Number of Posts	393157	3016144
Number of Users	2198	4124

- Collection: Take all tweets from 1 month period after the user self-declared that they have been diagnosed with depression (self-declared tweets were removed)
- Raw text of each Tweets, User id
- Binary Classification / Detection Task
- Post-level / User-level
- Performance Metrics :
    - Accuracy, Precision, Recall, F1-score

Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution
Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu
Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence Main track. Pages 3838-3844. https://doi.org/10.24963/ijcai.2017/536

Performance on this dataset

Year	Title	Model	Performance (User-level)
2017	Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution	MDL (Multimodal Dictionary Learning)	F1 = 0.85
2019	Cooperative Multimodal Approach to Depression Detection in Twitter	COMMA indicator selection GRU + VGG-Net	F1 = 0.90
2020	SenseMood: Depression Detection on Social Media	BERT + CNN	F1 = 0.93
2021	DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media	Stacked_BiGRU + CNN-BiGRU_Att BERT-BART	F1 = 0.912

Problems : Mental health-related keywords

List_1 = Top 1000 words from TF-IDF of Depression Class
List_2 = Top 1000 words from TF-IDF of Control Class
List_1 - List_2

['**depression**', '**health**', '**mental**', '**anxiety**', 'com', 'article', '**disorder**',
'**askdoctorforfree**', '**illness**', '**bipolar**', '**doctor**', '**treatment**', 'advice', 'foto', 'weight',
'tip', '**suffer**', '**diagnose**', 'cancer', 'omfg', '**drug**', 'struggle', 'friday', 'yay',
'husband', 'healthy', 'scar', '**depress**', 'teacher', '**suicide**', 'anonymous', 'org', 'luke',
'**symptom**', 'tattoo', 'raise', '**healthcarechange**', '**hospital**', 'sunday', 'justin', 'harry',
'cure', 'cuddle', 'period', 'cause', 'normal', 'upset', 'cameron', 'judge', '**therapy**',
'alex', 'student', 'gorgeous', 'yr', '**medical**', 'bully', '**disease**', 'stats', 'hopefully',
'advance', 'saturday', 'xxx', 'insurance', 'ifb', 'inspire', 'monday', 'injury', 'horrible',
'community', 'terrible', 'bae', 'product', 'stomach', 'difference', '**ptsd**', 'conversation',
'concert', 'tour', 'complete', 'spread', 'sort', 'public', 'donate', 'email', 'adorable',
'**psychology**', 'five', 'recently', 'loss', 'headache', 'due', 'breathe', 'diet', 'include',
'cream', 'themselves', 'canada', 'natural', 'pregnant', 'manage', 'currently', 'heal',
'emotional', 'con', '**meds**', 'non', 'male', 'direction', 'lately', 'fave', 'pill',
'snapchat', 'shame', 'softly', 'four', 'bother', 'panic', 'gain', 'jack', '**adhd**', '**stigma**',
'hero', 'london', 'instagram', 'scary', 'common', 'spinal', 'fill', 'btw', 'police',
'awkward', 'reality', 'butt', 'acne', 'abuse', 'tumblr', 'awareness', 'acid', 'nose',
'practice', 'area', 'weather', 'dr', 'putt', 'surgery', 'congrats', 'whenever', 'thumb',
'lunch', 'cousin', 'affect', 'cord', 'overcome', '**medication**', 'effect', 'result', 'bus',
'crap', 'makeup', 'ear', 'benefit', 'yoga', 'shall', 'aww']

Problems

- High chance of model just memorizing the keywords
  - How would the model perform after removing the related words ?
  - How to avoid memorizing the keywords while maintaining the model's performance ?
- Understanding / Explainability
  - Why does the model classify a specific user/post as Depression ?
  - What are the language patterns of depression detected by the model ?
- Many Posts per one user, are ALL of them important ? (for user-level)
  - How to select only the important parts ?
- Class imbalance : Control >> Depression
  - How to make the model more robust to imbalance data ?
- etc.

>> Make reasonable assumptions as needed.

Related Works

Depression Detection in Social Media

Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution : https://doi.org/10.24963/ijcai.2017/536

MGL-CNN: A Hierarchical Posts Representations Model for Identifying Depressed Individuals in Online Forums : https://ieeexplore.ieee.org/document/8998086

Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study : https://arxiv.org/abs/2007.00228

DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media : https://arxiv.org/abs/2105.10878

SenseMood: Depression Detection on Social Media : https://dl.acm.org/doi/abs/10.1145/3372278.3391932

Explainabitilty

Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7298178/

Removing Keywords

Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP : https://aclanthology.org/W18-4102

Selective Masking in MLM

MASKER: Masked Keyword Regularization for Reliable Text Classification : https://arxiv.org/abs/2012.09392

Train No Evil: Selective Masking for Task-Guided Pre-Training : https://arxiv.org/abs/2004.09733

Knowledge Enhanced Masked Language Model for Stance Detection : https://aclanthology.org/2021.naacl-main.376