Tasks and Applications in NLP

He He

NEW YORK UNIVERSITY

February 21, 2023

Logistics

- Feb 28 lecture will be pre-recorded.
- Section will be in-person, starting at 4:55pm.
    - Review and Q&A about the lecture recording.
    - Lab material.

Table of Contents

Plan for today

- So far, we have viewed NLP tasks in a somewhat abstract way (classification, sequence generation).
- The actual tasks are much richer, each comes with its unique challenges.
- **Goal of today**: get a sense of what problems people are working on in NLP and maybe find your own problem!
- **Section**: where to find datasets and how to use them

Two categorizations of tasks

By **purpose**:

- Capabilities: test key abilities (linguistic, social, cultural, etc.) of language understanding
  e.g., parts-of-speech tagging, parsing, commonsense
- Application: a use case with potential products in mind
  e.g., machine translation, question answering
- NLP + X: new dimensions of NLP
  e.g., multilingual, multimodal, social NLP etc.

Two categorizations of tasks

By **purpose**:

- Capabilities: test key abilities (linguistic, social, cultural, etc.) of language understanding
  e.g., parts-of-speech tagging, parsing, commonsense
- Application: a use case with potential products in mind
  e.g., machine translation, question answering
- NLP + X: new dimensions of NLP
  e.g., multilingual, multimodal, social NLP etc.

By **modeling**:

- Classifcation: output is a categorical variable
- Structured prediction: output is a chain, a tree, a graph
- Generation: output is free-form text

Table of Contents

Basic text processing

Stanford CoreNLP

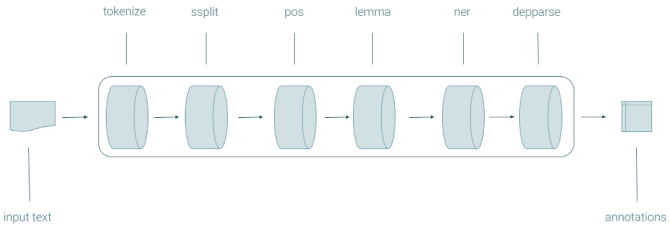


Figure: https://stanfordnlp.github.io/CoreNLP/

- Intermediate steps of a pipeline system
- Used by downstream models that are more directly connected to an application
- E.g., tokenization $\longrightarrow$ topic models

Parts-of-speech tagging

Assign each token a part-of-speech tag:

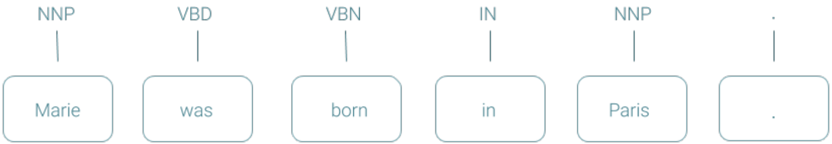


Figure: https://stanfordnlp.github.io/CoreNLP/

What is needed to perform this task well?

Parts-of-speech tagging

Assign each token a part-of-speech tag:

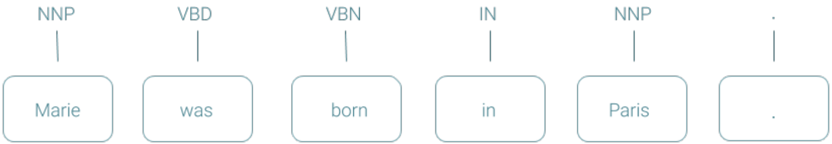


Figure: https://stanfordnlp.github.io/CoreNLP/

What is needed to perform this task well?

- Memorize possible tags for each word
- Model short range context

What can you do with the output of this task?

Parts-of-speech tagging

Assign each token a part-of-speech tag:

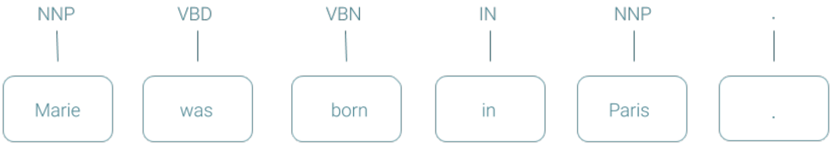


Figure: https://stanfordnlp.github.io/CoreNLP/

What is needed to perform this task well?

- Memorize possible tags for each word
- Model short range context

What can you do with the output of this task?

Named entity recognition

$\underbrace{\text{New York University}}_{\textit{org}}$ is a private research university based in $\underbrace{\text{New York City}}_{\textit{loc}}$. It is founded in $\underbrace{\text{1831}}_{\textit{year}}$ by $\underbrace{\text{Albert Gallatin}}_{\textit{people}}$.

Named entity recognition

$\underbrace{\text{New York University}}_{org}$ is a private research university based in $\underbrace{\text{New York City}}_{loc}$. It is founded in $\underbrace{\text{1831}}_{year}$ by $\underbrace{\text{Albert Gallatin}}_{people}$.

$\underbrace{\text{CT of the maxillofacial area}}_{test}$ showed no facial $\underbrace{\text{bone fracture}}_{symptom}$.

What is the challenge in this task?

Named entity recognition

$\underbrace{\text{New York University}}_{\textit{org}}$ is a private research university based in $\underbrace{\text{New York City}}_{\textit{loc}}$. It is founded in $\underbrace{\text{1831}}_{\textit{year}}$ by $\underbrace{\text{Albert Gallatin}}_{\textit{people}}$.

$\underbrace{\text{CT of the maxillofacial area}}_{\textit{test}}$ showed no facial $\underbrace{\text{bone fracture}}_{\textit{symptom}}$.

What is the challenge in this task?

- Variations of references to an entity (NYU, New York Uni)
- Ambiguity (Washington: state or people?)
    - Related task: entity linking (multiple people can be named Washington)

Named entity recognition

$\underbrace{\text{New York University}}_{org}$ is a private research university based in $\underbrace{\text{New York City}}_{loc}$. It is founded in $\underbrace{\text{1831}}_{year}$ by $\underbrace{\text{Albert Gallatin}}_{people}$.

$\underbrace{\text{CT of the maxillofacial area}}_{test}$ showed no facial $\underbrace{\text{bone fracture}}_{symptom}$.

What is the challenge in this task?

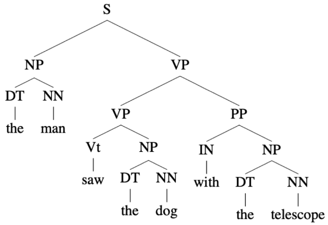
- Variations of references to an entity (NYU, New York Uni)
- Ambiguity (Washington: state or people?)
    - Related task: entity linking (multiple people can be named Washington)

Useful for information extraction or knowledge base construction

Parsing

Syntactic structures of a sentence

- Constituents: small components in a sentence that compose into larger ones

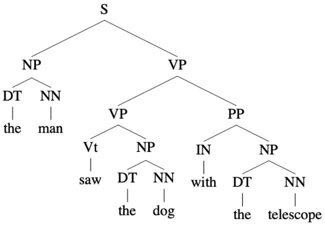


context free grammars

Parsing

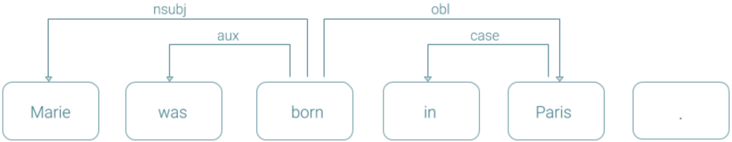
Syntactic structures of a sentence

- Constituents: small components in a sentence that compose into larger ones



context free grammars

- Dependencies: relations between words (modify, arguments of etc.)



Parsing

What are the challenges?

Parsing

What are the challenges?

- Design and annotate sentences with parse trees
- Parsing algorithm: find the highest scoring tree out of all possible trees
- Multilingual support

Parsing

What are the challenges?

- Design and annotate sentences with parse trees
- Parsing algorithm: find the highest scoring tree out of all possible trees
- Multilingual support

Why do we need parsing?

- A model that understands a sentence must understand its structure (even if not explicitly)
- More generally, it’s a study about compositionality (which is key to language understanding).

Coreference resolution

John had a great evening meeting with his high school friends.

What are the challenges?

Coreference resolution

John had a great evening meeting with his high school friends.

What are the challenges?

- Sometimes there're surface cues, othertimes it requires semantic understanding Easy Victories and Uphill Battles in Coreference Resolution [Durret and Klein, 2013]
- Commonsense reasoning (Winograd schema challenge)

  The city councilmen refused the demonstrators a permit because they feared violence.

Commonsense reasoning

Motivation: many tasks requires commonsense knowledge. Can we construct a separate test for it?

Commonsense reasoning

Motivation: many tasks requires commonsense knowledge. Can we construct a separate test for it?

Which one is the most likely continuation? (example from Hellaswag [Zellers et al., 2019])

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath.

A rinses the bucket off with soap and blow dry the dog's head.

B uses a hose to keep it from getting soapy.

C gets the dog wet, then it runs away again.

D gets into a bath tub with the dog.

Table of Contents

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

What is the use case?

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

What is the use case? content moderation

What are the challenges?

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

What is the use case? content moderation

What are the challenges?

- Toxicity may need to be interpreted in context [Pavlopoulos et al., 2020]
- Dataset biases (section)

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

What is the use case? content moderation

What are the challenges?

- Toxicity may need to be interpreted in context [Pavlopoulos et al., 2020]
    - Hi Gadget, interpreted in what manner? Flaming gays? Or Burn a gay?
- Dataset biases (section)

Toxicity classification

The profoundly stupid have spoken.	toxic
The president makes himself an easy target.	okay

What is the use case? content moderation

What are the challenges?

- Toxicity may need to be interpreted in context [Pavlopoulos et al., 2020]
  - Hmmm. The flame on top of the gay pride emblem can probably be interpreted in a manner that I did not consider. Perhaps one icon on each end using?
  - Hi Gadget, interpreted in what manner? Flaming gays? Or Burn a gay?
- Dataset biases (section)

Question answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure: SQuAD

Reading comprehension (close-book QA):

- Input: document and question
- Output: start and end indices of the answer span

What are the challenges?

Question answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure: SQuAD

Reading comprehension (close-book QA):

Input: document and question

Output: start and end indices of the answer span

What are the challenges?

- Long documents (see long text QA)
- Unanswerable questions (see SQuAD 2.0)

Question answering

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Figure: Natural questions

Open-domain question answering:

Input: question

Output: answer (in text)

What are the challenges?

Question answering

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Figure: Natural questions

Open-domain question answering:

Input: question

Output: answer (in text)

What are the challenges?

- Retrieval
- Evaluation (see equivalent answers)
- Presupposition (see Kim et al., 2021)
  What is the stock symbol for mars candy?

Summarization

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Figure: XSum

Abstractive summarization:

Input: document (e.g., a news article)

Output: summary (in text)

Extractive summarization:

Input: document

Output: k sentences from the document

What are the challenges?

Summarization

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Figure: XSum

Abstractive summarization:

Input: document (e.g., a news article)

Output: summary (in text)

Extractive summarization:

Input: document

Output: k sentences from the document

What are the challenges?

- Evaluation: what is a good summary?
- Faithfulness (see Durmus et al., 2020)

Semantic parsing

Natural language to formal language:

- Input: text (e.g., question, instruction)
- Output: logical form (DSL, e.g., SQL) $\longrightarrow$ execute to get result

Complex question What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL

```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

Figure: Spider

What are the use cases?

Semantic parsing

Natural language to formal language:

- Input: text (e.g., question, instruction)
- Output: logical form (DSL, e.g., SQL) $\longrightarrow$ execute to get result

Complex question What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL

```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

Figure: Spider

What are the use cases?

- Interface with a database, interpreter (shell, python)
- More generally, interact with a computer

Categorization of tasks by modeling

Classification: text $\rightarrow \{1, \ldots, K\}$

- E.g., Toxic classification, natural language inference, multiple choice QA

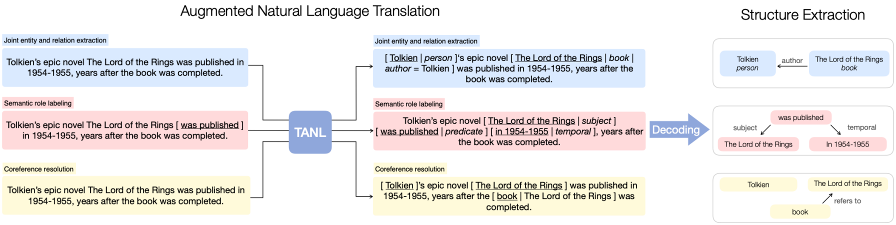
Structured prediction:

- Sequence labeling: $\mathcal{V}_{\text{in}}^{n} \rightarrow \mathcal{V}_{\text{out}}^{n}$
  - E.g., POS tagging, NER (using the BIO annotation), close-book QA
- Parsing: $\mathcal{V}_{\text{in}}^{n} \rightarrow$ tree
  - E.g., constituent, dependency, semantic parsing

Categorization of tasks by modeling

Generation: $\mathcal{V}_{\text{in}}^n \to \mathcal{V}_{\text{out}}^m$

- Classification: $m = 0, \mathcal{V}_{\text{out}} = \{1, \ldots, K\}$
- Structured prediction with linearized annotation



- Sequence to sequence, e.g., machine translation, summarization, text-to-code

The most general format (pros and cons?)

Table of Contents

Structured prediction

Exact match: unit of comparison is the whole structure

- output is correct only if it is exactly the same as the reference

Structured prediction

Exact match: unit of comparison is the whole structure

- output is correct only if it is exactly the same as the reference

How do we account for partial correct answers?

F1: unit of comparison is components of the structure

- Define components of the output
- Compute overlap of components in terms of F1 between each predicted structure and its reference
- Average the F1 score over all examples

Structured prediction

Exact match: unit of comparison is the whole structure

- output is correct only if it is exactly the same as the reference

How do we account for partial correct answers?

F1: unit of comparison is components of the structure

- Define components of the output
- Compute overlap of components in terms of F1 between each predicted structure and its reference
- Average the F1 score over all examples

Example: reading comprehension

- predicted = skilled workers = $\{\text{skilled}, \text{workers}\}$
- reference = an increase in skilled workers = $\{\text{skilled}, \text{workers}, \text{an}, \text{increase}, \text{in}\}$
- precision =
- recall =

Generation

Task: given the reference(s) of each source sentence, evaluate the quality of the generated sequences.

Reference 1	It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2	It is the guiding principle which guarantees the military forces always being under the command of the Party.
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party.
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party direct.

Main idea: good generations should have high overlap with the reference.

BLEU: n-gram precision

First try: n-gram precision (x: input, c: candidate, r: reference)

$$p_n = \frac{\sum_{(x,c,r)} \sum_{s \in \text{n-gram}(c)} \mathbb{I}\left[s \text{ in } r\right]}{\sum_{(x,c,r)} \sum_{s \in \text{n-gram}(c)} \mathbb{I}\left[s \text{ in } c\right]} = \frac{\text{\# n-grams in both cand and ref}}{\text{\# n-grams in cand}}$$

BLEU: n-gram precision

First try: n-gram precision (x: input, c: candidate, r: reference)

$$p_n = \frac{\sum_{(x,c,r)} \sum_{s \in \text{n-gram}(c)} \mathbb{I}[s \text{ in } r]}{\sum_{(x,c,r)} \sum_{s \in \text{n-gram}(c)} \mathbb{I}[s \text{ in } c]} = \frac{\text{\# n-grams in both cand and ref}}{\text{\# n-grams in cand}}$$

Problem: can match only a few words in the reference(s)

Candidate	the the the the the the the
Reference 1	The cat is on the mat
Reference 2	There is a cat on the mat

unigram precision = ?

Solution: clip counts to maximum count in the reference(s)

BLEU: combine n-gram precision

Compute n-gram precision for each n (typically up to 4)

Then, we need to combine the n-gram precisions.

Average? Problem: precision decreases roughly exponentially with n.

BLEU: combine n-gram precision

Compute n-gram precision for each n (typically up to 4)

Then, we need to combine the n-gram precisions.

Average? Problem: precision decreases roughly exponentially with n.

Solution: geometric mean (when $w_n = 1/n$)

$$\exp\left(\sum_{i=1}^{n} w_n \log p_n\right)$$

BLEU: brevity penalty

Problem with precision: “One who does nothing also does nothing wrong”

Candidate	of the
Reference 1	It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 2	It is the practical guide for the army always to heed the directions of the party.

Why not use recall?

BLEU: brevity penalty

A good translation must match the reference in:

word choice	captured by precision
word order	capture by n-gram
length	?

candidate length $C = \sum_{(x,c,r)} \text{len}(c)$

reference length $R = \sum_{(x,c,r)} \arg\min_{a \in \{\text{len}(r_1),\ldots,\text{len}(r_k)\}} |a - \text{len}(c)|$

- Use the reference whose length is closest to the candidate

Brevity penalty $BP = \begin{cases} 1 & \text{if } c \geq r & \text{no penalty} \\ e^{1-R/C} & \text{if } c < r & \text{downweight score} \end{cases}$

BLEU

Putting everything together:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\log \text{BLEU} = \min(1 - \frac{R}{C}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

- Both precision and the brevity penalty are computed at the *corpus level*.
- Need smoothing for sentence-level BLEU.
- Good correlation with human evaluation for MT (typically $n = 4$).

ROUGE

Task: given a candidate summary and a set of reference summaries, evaluate the quality of the candidate.

ROUGE-n: n-gram recall

- Encourage content coverage

ROUGE-L: measures longest common subsequence between a candidate and a reference (doesn't require consecutive match.)

- Precision $= LCS(c, r)/\text{len}(c)$
- Recall $= LCS(c, r)/\text{len}(r)$
- F-measure $= \frac{(1+\beta^2)RR}{R+\beta^2 P}$

Automatic evaluation metrics for generation

n-gram matching metrics (e.g. BLEU, ROUGE)

- Measures exact match with reference; interpretable.
- Do not consider semantics.

Embedding-based metrics (e.g. BERTScore, MAUVE)

- Measures similarity to the reference in an embedding space.
- Captures synonyms and simple paraphrases.

However, we also want to measure

- Is the generation correct? e.g. faithfulness (summarization), adequacy (MT).
- Open-ended generation: is the story/dialogue interesting, informative, engaging?
- So **human evaluation** is still needed.

Table of Contents

Common types of projects

Find a nail: identify a problem/domain that you are excited about and try to solve it using whatever method that works

Automated Machine Transcriptions for Handwritten Historical Documents from NYPL Digital Collections

João Galinho
New York University
joao.galinho@nyu.edu

Diogo Vieira
New York University
diogo.vieira@nyu.edu

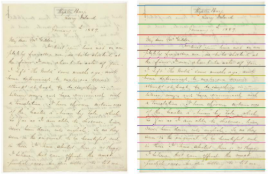


Figure 2: Original handwritten document (left) and corresponding line segmentation output (right).

Likely to succeed if:

- You know a domain and its challenges very well
- You have access to (high-quality, large) data (important!)
- You have a reliable way to evaluate the result

Common types of projects

Find a hammer: identify a method that you are excited about and try to improve or extend it on its common use cases

Studying the Effect of Generalized Entropy Regularization on Hierarchical Story Generation

Anya Trivedi aht324@nyu.edu
Vishal Kumar vk2161@nyu.edu
Mahima Gaur mg6827@nyu.edu

Thus, the loss function to be optimized is described as

$$L(\theta) + \beta R(\theta) \tag{1}$$

where

$$\begin{aligned} L(\theta) &= \mathrm{KL}(\tilde{p} || p_\theta) & (2) \\ &= H(\tilde{p}, p_\theta) - H(\tilde{p}) & (3) \end{aligned}$$

Likely to succeed if:

- You know a method and its variants/extensions well
- You have identified a weakness (e.g., efficiency, reliability, problem-specific challenges)

Common types of projects

Study a nail or hammer: analyze common methods and their applications

Evaluating Prompts Across Multiple Choice Tasks In a Zero-Shot Setting

Gabriel Orlanski
go533@nyu.edu

		ANLI R1	ANLI R2	ANLI R3	AQuA	CB	Craigslist	RTE	WiC	Rank
	No Prompt	34.15	33.35	33.42	26.77	24.11	16.83	59.57	50.24	46.25
Unseen Prompts	ANLI	**37.60**	**34.70**	**34.08**	25.95	**32.14**	21.44	64.62	50.16	24.50
	AQuA	36.10	33.40	35.42	**17.32**	33.93	23.45	71.12	51.57	18.25
	COPA	39.30	34.40	34.00	20.47	26.79	16.58	69.31	50.63	21.25
	Craigslist	31.40	31.30	32.83	25.79	8.04	**26.72**	49.82	50.16	71.25
	MathQA	37.30	33.50	34.25	19.29	26.79	16.25	73.29	51.10	24.50
	RTE	36.10	33.20	33.58	22.05	23.21	20.27	**61.37**	50.47	43.25
	SemEval2010	33.10	32.00	32.58	27.56	14.29	25.63	55.23	50.47	66.50
	WiC	31.75	33.45	32.33	26.57	13.39	18.01	55.05	**50.47**	64.25
Training Prompts	AppReviews	34.20	33.10	33.62	27.17	19.64	33.17	61.55	50.31	33.50
	IMDB	33.00	32.20	33.08	26.38	12.50	14.57	55.23	50.16	71.25
	Yelp	33.25	32.35	33.04	26.77	12.50	24.29	62.27	51.57	41.75

Table 2: Median Accuracy when using modified prompts for cross task zero-shot evaluation. **Bolded** entries are prompts for the original task. Green Cells and Red Cells are the best and worst performing tasks for a column respectively. Rank is the median rank of prompts from this task out of 95 total prompts. ANLI and CB both use the same prompts for their original task prompts per PromptSource. Some tasks are left out for clarity. The full table can be found in Table 6.

Likely to succeed if:

- You have an interesting question to ask
- You are good at running large scale experiments

Project proposal

Before submitting the proposal:

- Form groups and identify a rough topic of interest
- Literature survey
- Get all resource ready (data, codebase, machines)

Write the proposal:

- Overview
  - What problem are you going to work on?
  - What are the challenges?
  - What's your solution?
- Project plan
  - What do you plan to do (experiments, data, model)
  - How do you evaluate success?