CMPT 413/713: Natural Language Processing

Project tips and analyzing your results

Spring 2025

2025-03-12

Project Milestone

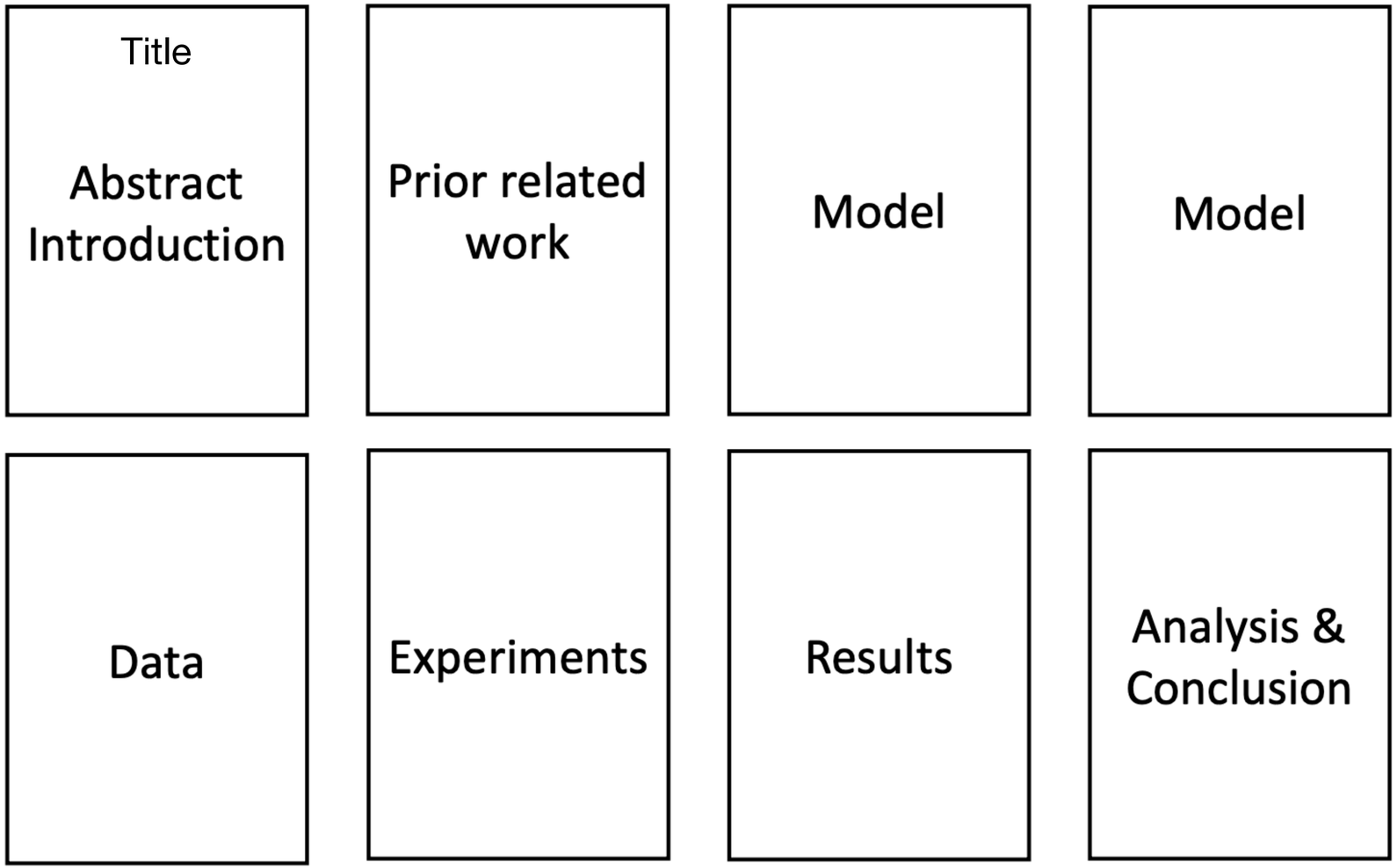
- Project Milestone due Thursday 3/20
- PDF (3-6 pages) in the style of a conference (e.g. ACL/EMNLP) submission
  - https://2020.emnlp.org/files/emnlp2020-templates.zip
- Milestone should include:
  - **Title and Abstract** - summary of what you are working on
  - **Introduction** - motivate the problem, describe your goals, and highlight your findings
  - **Prior Work** - what have others done in this area?
  - **Approach** - details on your main approach and baselines. Be specific. Make clear what part is original, what code you are writing yourself, what code you are using
  - **Experiment** - describe dataset, evaluation metrics, what experiments you plan to run, any results you have so far. Also provide training details, training times, etc.
  - **Future Work** - what is your plan for the rest of the project
  - **Reference** - provide references using BibTex
- Milestone will be graded based on **progress** and **writing quality**

Final project report

8 pages not including references

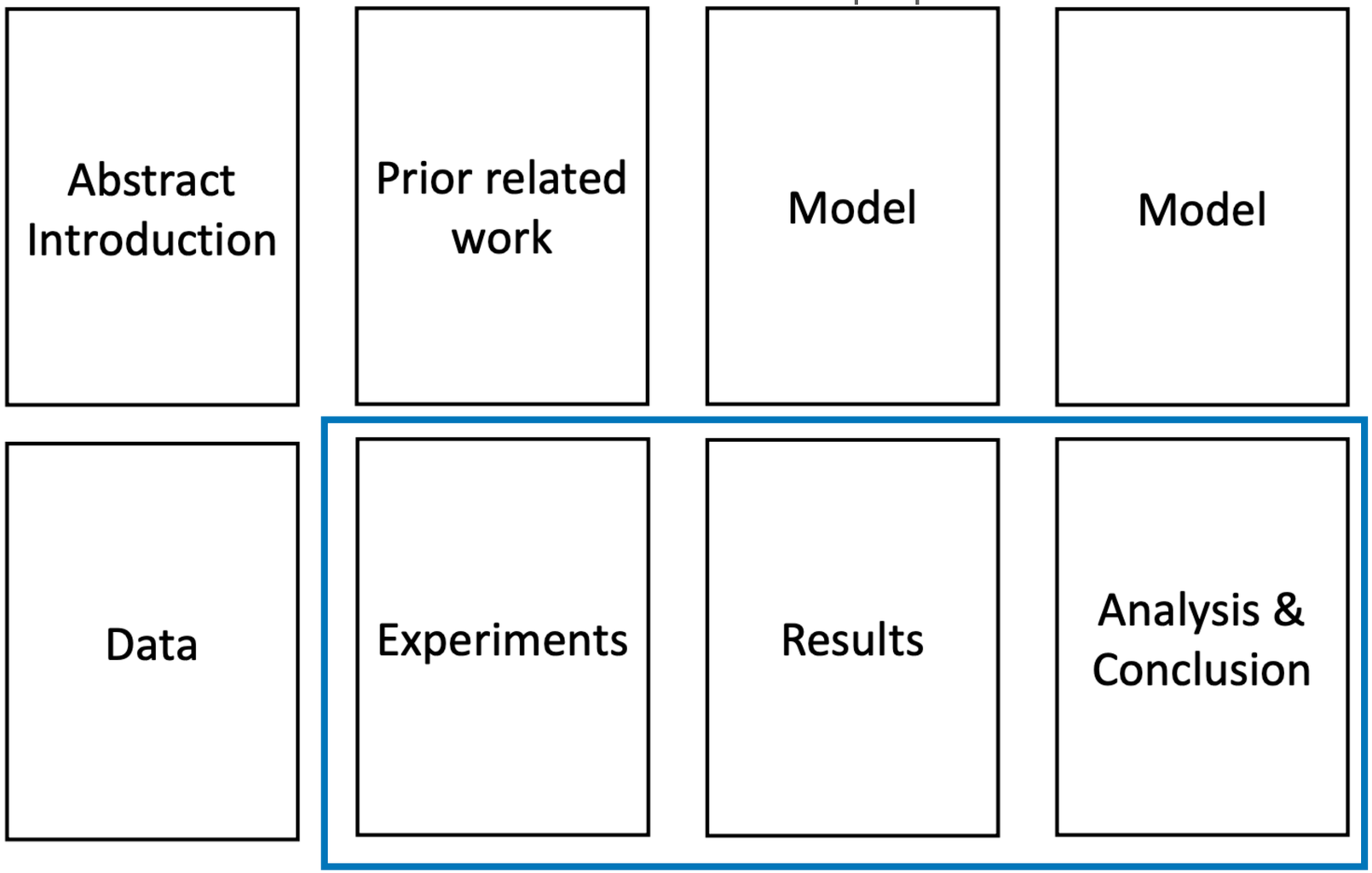
Model or what approach you are taking

Rough page layout: your report can have different number of pages for each section



Milestone report as a draft for final project report

We expect you to have around 3-6 pages, can be more (but < 8) if you have done lots of work since the proposal



Have **statistics**/ **analysis** of your data, and show **examples**!

Not expected to be complete:

- preliminary experiments
- difficulties
- what is still to be done

Include References

Why a page limit?

- Good writing is concise and clear
- More words does not mean more information or higher quality
- Trim words you don’t need.
- But I have a lot of stuff to say!
  - References don’t count toward the page limit
  - You can have appendices if you really want to share (the appendices will not be graded)

Project Milestone

- Build on what you have done for the proposal
- Include progress and initial results
- Flesh out description of your approach
- Include figures for your model
- Include dataset examples and statistics
- Have plan for experiments and analysis you will do
- Include references
- Make it clear what you are implementing vs what part you are building on top of existing libraries / codebases / homework

What you should have from the Project Proposal:

- What **task** are you addressing? What is the **input / output**? Why is it interesting?
- What specific aspects will your project be on?
  - Re-implement paper? Compare different methods? Analysis?
- What have others (**prior work**) done to address the same problem?
- What **data** do you plan to use?
  - Preliminary statistics for your data (number of sentences, tokens, etc)
- What is the specific method or methods you will use to address the task?
  - **What will you implement by yourself vs what existing code will you use?**
  - What **compute resources** do you plan to use?
- How do you plan to evaluate?
  - Data splits?
  - What **metrics**?
  - What experiments will you run to **compare** different variations / different approaches?
- **Timeline and work breakdown**
  - What do you plan to have by the milestone? The end of the term?
  - Who will work on what?

Project Milestone

Build on what you have done for the proposal

- **Progress**: What have you achieved so far? Are there any issues you encountered?
- **Task / Problem State:**
  - Clearly state the input and output
- **Related Work:** How have others approached this problem?
- **Approach:** Flesh out and update your approach based on what you have learned.
  - Make it clear what you are implementing vs what part you are building on top of existing libraries / codebases / homework
    - Specify what you have implemented and what you still have to implement.
  - Focus on describing how you are using a particular model for your task. For instance, if you are using a RNN for text classification, explain how the input is fed into the RNN, and how the RNN is used to make predictions. What is the training loss that you will use?
  - Make figures (if appropriate) that illustrate your approach.
  - Using equations (if appropriate) with clear mathematical style to explain your model(s).

Project Milestone

Build on what you have done for the proposal

- **Data:** At this point, you should know what data you will be using. Describe the data and provide some statistics and examples from the data.
- **Experiments and results:** At this point, you should have a clear idea of what experiments you will perform and/or some preliminary results.
  - Provide a list of experiments you have will perform. Describe what you expect the experiments to reveal, or what is uncertain about the potential outcomes.
  - Provide a summary of your preliminary results, and describe remaining results that you plan to produce
  - Have a empty table with rows and columns
- **Timeline / Plans for remaining tasks**: Present an updated timeline of the planned tasks/goals.
  - Clearly state what you plan to complete by the final report.
  - If you are working in a group, please also state the contribution of each team member
- **References:** Many of you already had references in the proposal. For the milestone, this is required.

Proposal Observations

- **Task**
  - There are standard NLP tasks: sentiment analysis, question answering
  - Tasks are defined by their input and output - so make sure you clearly specify that
  - Simple and effective way to do it is via an example

- Summarize **relevant work**
  - Summarize what did they do? What was the key findings?
  - (this requires larger changes than just replacing “we” with “they”)
  - Avoid copying text from another paper verbatim (if you copy, it is plagiarism).
  - Describe the approach in your own words. Use different sentence structures.

Task / method illustration example

Clear input

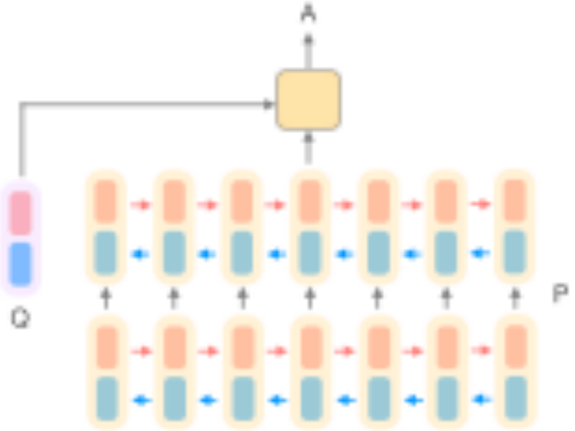
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever

Document Reader

Clear output

833,500



Reading Wikipedia to Answer Open-Domain Questions [Chen et al, 2017]

Proposal Observations

- **Approach / Method / Model**
  - This is what you proposed to do to solve the task
    - Describe how you will use a particular method for your task
    - It's great that you are using RNN, LSTM, GRU, and that transformers uses self-attention, and know the differences between the models.
    - But how will you actually use it for your task?
  - Typically in NLP we use **neural network models**, so often it is
    - a description of the model architecture with information on
    - how the **input** is fed in
    - how the **prediction** is made
    - how the model is trained (**optimizer and loss function**)

Method figure example

Figure showing key elements

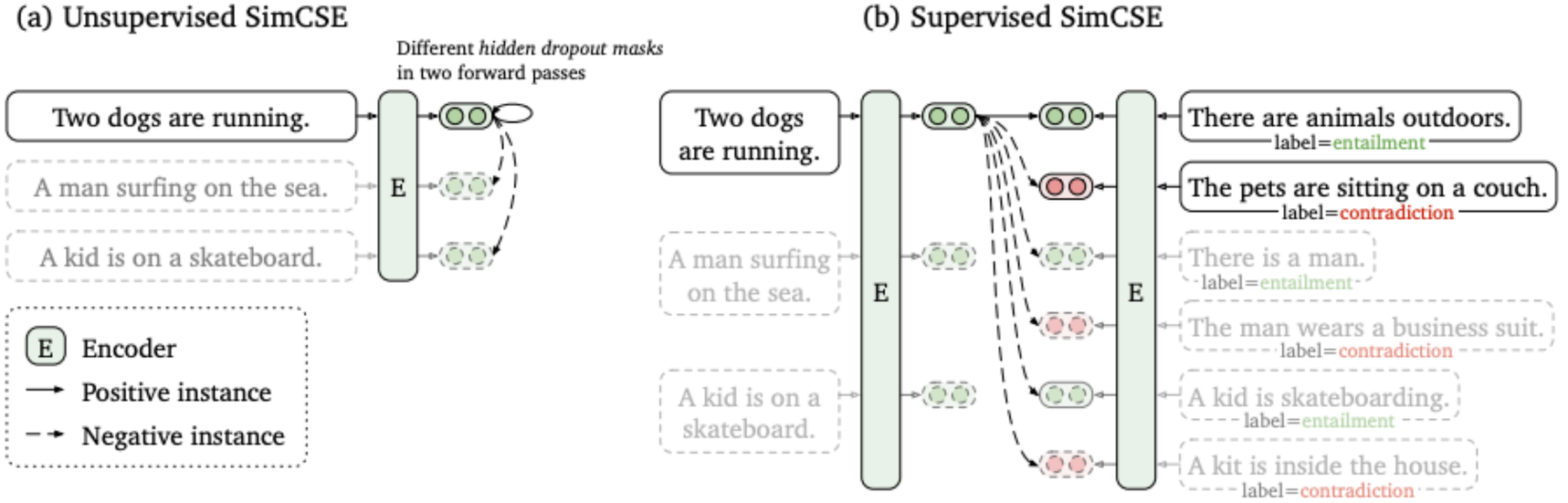


Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

Training objective (loss function)

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}}$$

SimCSE: Simple Contrastive Learning of Sentence Embeddings [Gao et al, 2022]

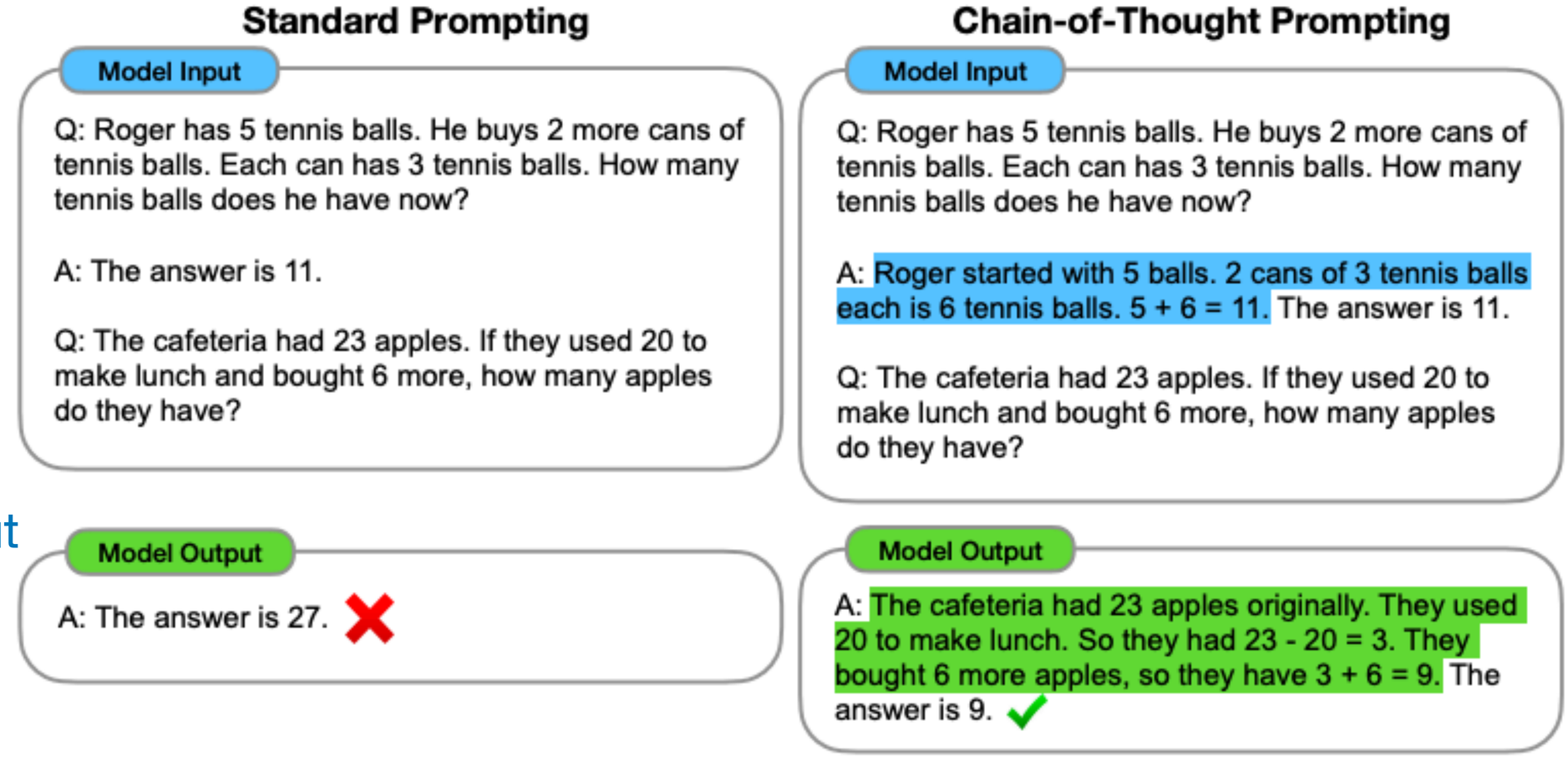
Proposal Observations

- If you are doing **prompt engineering** with large language models
  - Clearly describe the prompts you are trying out
  - If you are giving the LLM examples, make sure to clearly specify how the examples are specified
  - Clearly specify how you will interpret the LLM output.
    - Do you see errors from the LLM?
    - How do you handle unexpected responses from the LLM
  - Be systematic

Prompting example

Clear input

Clear difference in prompting



Clear output

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Wei et al, 2022]

Proposal Observations

- Experiments
  - The quality of a model / approach is measured by **evaluation metrics**
  - For machine learning models, you will train and evaluate them with **data**. Specify clearly your data and statistics about your data.
  - Typically you want have a set of **comparisons** (often informed by some hypothesis)
    - There can be **variations** on your model depending on hyper-parameters, input encoding, how it is trained (training data, optimizer, loss function), etc.
      - Hypothesis: Adam converges faster than SGD. -> Experiment: Train with Adam and SGD and compare.
      - Hypothesis: Bidirectional RNNs gives higher performance than unidirectional models for text classification -> Experiment: Train Bi/Uni-dir models and compare.
    - You can also compare different models
  - Try to be **concrete** about your plans

Proposal Observations

- Use the correct format

- References
  - Use bibtex and \cite commands
  - Make sure that you cite the papers that introduced the method you are using (word2vec, BERT, LSTM, GRU, Transformer, etc).
  - Make sure that the references are properly included

- English
  - Use complete sentences
  - Proofread your report
  - Have friend (native/fluent English speaker) proofread your report

Timeline

- Make sure to allocate time for training and debugging the training of your model.
- In addition, you should allocate time for setting up the experiments (train classifier with some set of hyperparameters), evaluate the classifier, and analyzing the results.
- Plan for report writing and video making

Tips for good final projects

- Have a clear, well-defined hypothesis to be tested

  (++ novel/creative hypothesis)

- Conclusions and results should teach the reader something

- Meaningful tables, plots to display the key results

++ nice visualizations or interactive demos

++ novel/impressive engineering feat

++ good results

What to avoid

- All experiments run with prepackaged source - no extra code written for model/data processing
- Just ran model once or twice on the data and reported results (not much hyperparameter search done)
- A few standard graphs: loss curves, accuracy, without any analysis
- Results/Conclusion don't say much besides that it didn't work
  - Even if results are negative, you should analyze them!

Remember: Include analysis!

What makes for a good paper? (example paper)

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Task definition

Example with input and output

Text describing the problem

an example[4]: it consists of a passage p, a question q and an answer a, where the passage is a news article, the question is a cloze-style task, in which one of the article's bullet points has had one entity replaced by a placeholder, and the answer is this questioned entity. The goal is to infer the missing entity (answer a) from all the possible entities which appear in the passage. A news article is usually

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question

characters in " @placeholder " movies have gradually become more diverse

Answer

@entity6

Taken from the Reading Comprehension dataset introduced by Herman et al, 2015

Goal: identify entity that is goes where @placeholder goes

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Task definition

- Dataset is automatically generated
  - Use Google NLP pipeline and get entities and coreference chains
  - Cloze-style (fill in the blank) questions generated by taking sentence from passage and replacing entity reference with @placeholder
  - Data is anonymized (entities are just @entity#)

Example with input and output

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question

characters in " @placeholder " movies have gradually become more diverse

Answer

@entity6

Taken from the Reading Comprehension dataset introduced by Herman et al, 2015

Goal: identify entity that goes where @placeholder is

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Anonymization

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

Teaching machines to read and comprehend
https://arxiv.org/pdf/1506.03340.pdf (Hermann et al, NIPS 2015)

Approach

Two approaches presented

In this section, we describe two systems we implemented – a conventional entity-centric classifier and an end-to-end neural network. While Hermann

Features are described for entity-centric classifier

Describe problem again (with math symbols)

Given the (passage, question, answer) triple (p, q, a), $p = \{p_1, \ldots, p_m\}$ and $q = \{q_1, \ldots, q_l\}$ are sequences of tokens for the passage and question sentence, with q containing exactly one "@placeholder" token. The goal is to infer the correct entity $a \in p \cap E$ that the placeholder corresponds to, where E is the set of all abstract entity markers. Note that the correct answer entity must appear in the passage p.

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Model description

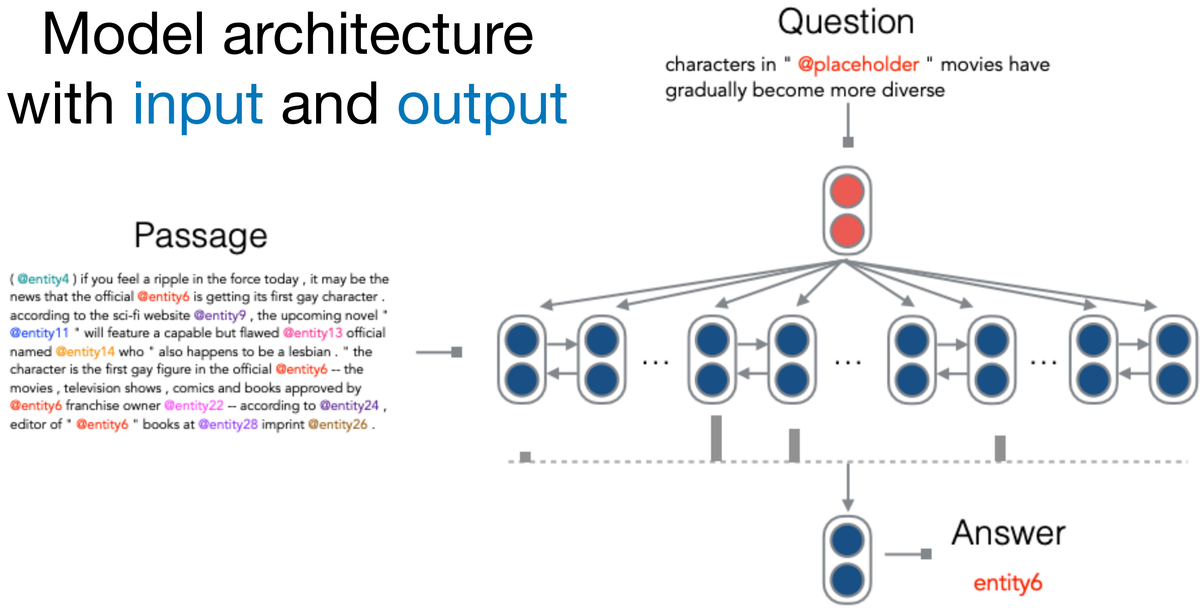
- Describe key parts including how it is hooked out to input and how prediction is made

Encoding: First, all the words are mapped to d-dimensional vectors via an embedding matrix $E \in \mathbb{R}^{d \times |\mathcal{V}|}$; therefore we have p: $\mathbf{p}_1, \ldots, \mathbf{p}_m \in R^d$ and $q : \mathbf{q}_1, \ldots, \mathbf{q}_l \in R^d$.

Prediction: Using the *output* vector $\mathbf{o}$, the system outputs the most likely answer using:

$$a = \arg\max_{a \in p \cap E} W_a^{\intercal} \mathbf{o} \qquad (4)$$

Model architecture with input and output



Symbols are connected

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Provide implementation and training details

- What external software / code was used
- Why was it used?

Training Details (for entity-centric model)

For training our conventional classifier, we use the implementation of *LambdaMART* (Wu et al., 2010) in the RankLib package.[5] We use this ranking algorithm since our problem is naturally a ranking problem and forests of boosted decision trees have been very successful lately (as seen, e.g., in many recent Kaggle competitions). We do not use all the features of *LambdaMART* since we are only scoring 1/0 loss on the first ranked proposal, rather than using an IR-style metric to score ranked results. We use Stanford's neural network dependency parser (Chen and Manning, 2014) to parse all our document and question text, and all other features can be extracted without additional tools.

Provide implementation and training details

- Vocabulary size, UNK handling
- Word embeddings
- Parameter initialization
- Dimensions (LSTM hidden state, word embeddings)
- Optimizer: type, learning rate, mini-batch size, dropout
- Number of epochs trained, how the “best” model was selected
- Compute resource uses (GPU type, runtime)

Training Details (for NN model)

For training our neural networks, we only keep the most frequent $|\mathcal{V}| = 50$k words (including entity and placeholder markers), and map all other words to an <*unk*> token. We choose word embedding size $d = 100$, and use the 100-dimensional pre-trained *GloVe* word embeddings (Pennington et al., 2014) for initialization. The attention and output parameters are initialized from a uniform distribution between $(-0.01, 0.01)$, and the LSTM weights are initialized from a Gaussian distribution $\mathcal{N}(0, 0.1)$.

We use hidden size $h = 128$ for *CNN* and 256 for *Daily Mail*. Optimization is carried out using vanilla stochastic gradient descent (SGD), with a fixed learning rate of 0.1. We sort all the examples by the length of its passage, and randomly sample a mini-batch of size 32 for each update. We also apply dropout with probability 0.2 to the embedding layer and gradient clipping when the norm of gradients exceeds 10.

All of our models are run on a single GPU (GeForce GTX TITAN X), with roughly a runtime of 6 hours per epoch for *CNN*, and 15 hours per epoch for *Daily Mail*. We run all the models up to 30 epochs and select the model that achieves the best accuracy on the development set.

Provide implementation and training details

Organize hyper parameters in a table if appropriate

Hyperparam	BERT	RoBERTa	ALBERT
Epochs	3, 10, 20	3	3
Learning rate	1e−5 − 5e−5	1e−5 − 3e−5	1e−5 − 3e−5
Learning rate schedule	warmup-linear	warmup-linear	warmup-linear
Warmup ratio	0.1	0.1	0.1
Batch size	16	16	16
Adam ϵ	1e−6	1e−6	1e−6
Adam β_1	0.9	0.9	0.9
Adam β_2	0.999	0.98	0.999
Adam bias correction	{True, False}	{True, False}	{True, False}
Dropout	0.1	0.1	–
Weight decay	0.01	0.1	–
Clipping gradient norm	1.0	–	1.0
Number of random seeds	25	25	25

On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines
https://arxiv.org/pdf/2006.04884v3.pdf (Mosbach et al, ICLR 2021)

Provide data statistics

- Statistics about what goes into train/val/test
- NLP specific statistics (number of tokens, sentences, unique words, etc)

	CNN	Daily Mail
# Train	380,298	879,450
# Dev	3,924	64,835
# Test	3,198	53,182
Passage: avg. tokens	761.8	813.1
Passage: avg. sentences	32.3	28.9
Question: avg. tokens	12.5	14.3
Avg. # entities	26.2	26.2

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Summarize your results in tables

Describe key findings in the text

(Hermann et al., 2015)

(Hill et al., 2016)

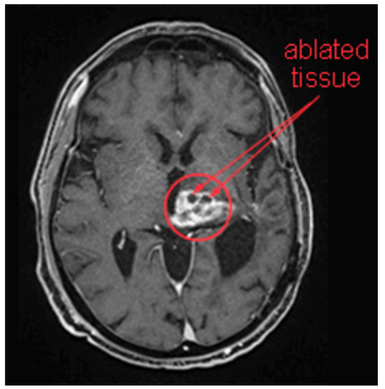
Model	CNN		Daily Mail	
	Dev	Test	Dev	Test
Frame-semantic model †	36.3	40.2	35.5	35.5
Word distance model †	50.5	50.9	56.4	55.5
Deep LSTM Reader †	55.0	57.0	63.3	62.2
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (window memory) ‡	58.0	60.6	N/A	N/A
MemNNs (window memory + self-sup.) ‡	63.4	66.8	N/A	N/A
MemNNs (ensemble) ‡	66.2*	69.4*	N/A	N/A
Ours: Classifier	67.1	67.9	69.1	68.3
Ours: Neural net	**72.4**	**72.4**	**76.9**	**75.8**

Ranking classifier with decision forests
Careful feature selection

NN Model is similar to Attentive Reader, compared to Attentive Reader
- Use bilinear attention instead of tanh layer
- Simpler model (only use weights contextual embeddings, don't classify over all vocabulary but only words that appear in passage)

Conduct ablation studies

to ablate = to remove



- If classical classifier (Naive Bayes, logistic regression, decision forest), do feature ablation to see how important is each feature

- If neural network with different components, ablate components.

- If loss is combination of terms, ablate loss terms.

Features	Accuracy
Full model	67.1
– whether e is in the passage	67.1
– whether e is in the question	67.0
– frequency of e	**63.7**
– position of e	65.9
– n-gram match	**60.5**
– word distance	65.4
– sentence co-occurrence	66.0
– dependency parse match	65.6

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Perform data analysis

- Manual analysis of 100 samples from val (dev) split
- Categorize into 6 different types (based on relationship of passage, question and answer)

No.	Category	(%)
1	Exact match	13
2	Paraphrasing	41
3	Partial clue	19
4	Multiple sentences	2
5	Coreference errors	8
6	Ambiguous / hard	17

- Answerable categories:
  - Exact match
  - Paragraphing
  - Partial clue
  - Multiple sentences
- Unanswerable categories:
  - Coreference error
    - The dataset was automatically generated
  - Ambiguous or very hard
    - Humans are not likely to be able to answer these

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
https://aclanthology.org/P16-1223.pdf (Chen et al, ACL 2016, Outstanding Paper)

Concrete examples for each category

Category	Question	Passage
Exact Match	*it 's clear @entity0 is leaning toward* **@placeholder** , says an expert who monitors @entity0	. . . @entity116 , who follows @entity0 's operations and propaganda closely , recently told @entity3 , *it 's clear @entity0 is leaning toward* **@entity60** in terms of doctrine , ideology and an emphasis on holding territory after operations . . . .
Paraphrase	**@placeholder** says he understands why @entity0 wo n't play at his tournament	. . . @entity0 called me personally to let me know that he would n't be playing here at @entity23 , " **@entity3** said on his @entity21 event 's website . . . .
Partial clue	a tv movie based on @entity2 's book **@placeholder** casts a @entity76 actor as @entity5	. . . to @entity12 @entity2 professed that his **@entity11** is not a religious book . . . .
Multiple sent.	he 's doing a his - and - her duet all by himself , @entity6 said of **@placeholder**	. . . we got some groundbreaking performances , here too , tonight , @entity6 said . we got **@entity17** , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . . . .
Coref. Error	rapper **@placeholder** " disgusted , " cancels upcoming show for @entity280	. . . with hip - hop star **@entity246** saying on @entity247 that he was canceling an upcoming show for the @entity249 . . . . (but @entity249 = @entity280 = SAEs)
Hard	pilot error and snow were reasons stated for **@placeholder** plane crash	. . . a small aircraft carrying **@entity5** , @entity6 and @entity7 the @entity12 @entity3 crashed a few miles from @entity9 , near @entity10 , @entity11 . . . .

Analysis leads to insights

- How many can the two models get correct for each category?
- Neural network is at close to optimal performance on this dataset!

(%)	Category	Classifier	Neural net
13	Exact match	13 (100.0%)	13 (100.0%)
41	Paraphrasing	32 (78.1%)	39 (95.1%)
19	Partial clue	14 (73.7%)	17 (89.5%)
2	Multiple sentences	1 (50.0%)	1 (50.0%)
8	Coreference errors	4 (50.0%)	3 (37.5%)
17	Ambiguous / hard	2 (11.8%)	1 (5.9%)
	All	66 (66.0%)	74 (74.0%)

Analyzing your results

Analyzing your results

- Understanding the output of your model
  - Look at the output and examine what it is getting right and what it is getting wrong
    - Provide examples
  - Conduct data / error analysis
  - Visualize
- Characterizing your model performance
  - Compare against other models
  - Compare against variations
    - Ablation studies

Provide qualitative examples of your model output

- Simple tip: color code correct and incorrect output

English-German translations

src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und *Miranda Kerr* lieben sich noch immer
best	Orlando Bloom und *Miranda Kerr* lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .
src	″ We ′ re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , ″ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht ” , sagte *Roger Dow* , CEO der U.S. Travel Association .
best	″ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit *unvereinbar* ist ″ , sagte *Roger Dow* , CEO der US - die .
base	″ Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ″ , sagte *Roger* **Cameron** , CEO der US - <unk> .

Effective Approaches to Attention-based Neural Machine Translation
https://arxiv.org/pdf/1508.04025.pdf (Luong et al, 2015)

Conduct data and error analysis

How to perform data and error analysis?

How to start?

- Take a manual subsample of the data
- Manually group and categorize them (like the paper we looked at)
  - You will need to figure out what these groups are!
- Compute some statistics
  - Overall percentage of categories
  - What is the performance on each category?

Adapted from slide by Graham Neubig

Conduct data and error analysis

How to perform data and error analysis?

General recipe

- Partition the performance of the validation set into different interpretable grouped based on pre-defined attributes.

  - Define attributes
  - Group test samples
  - Breakdown of performance

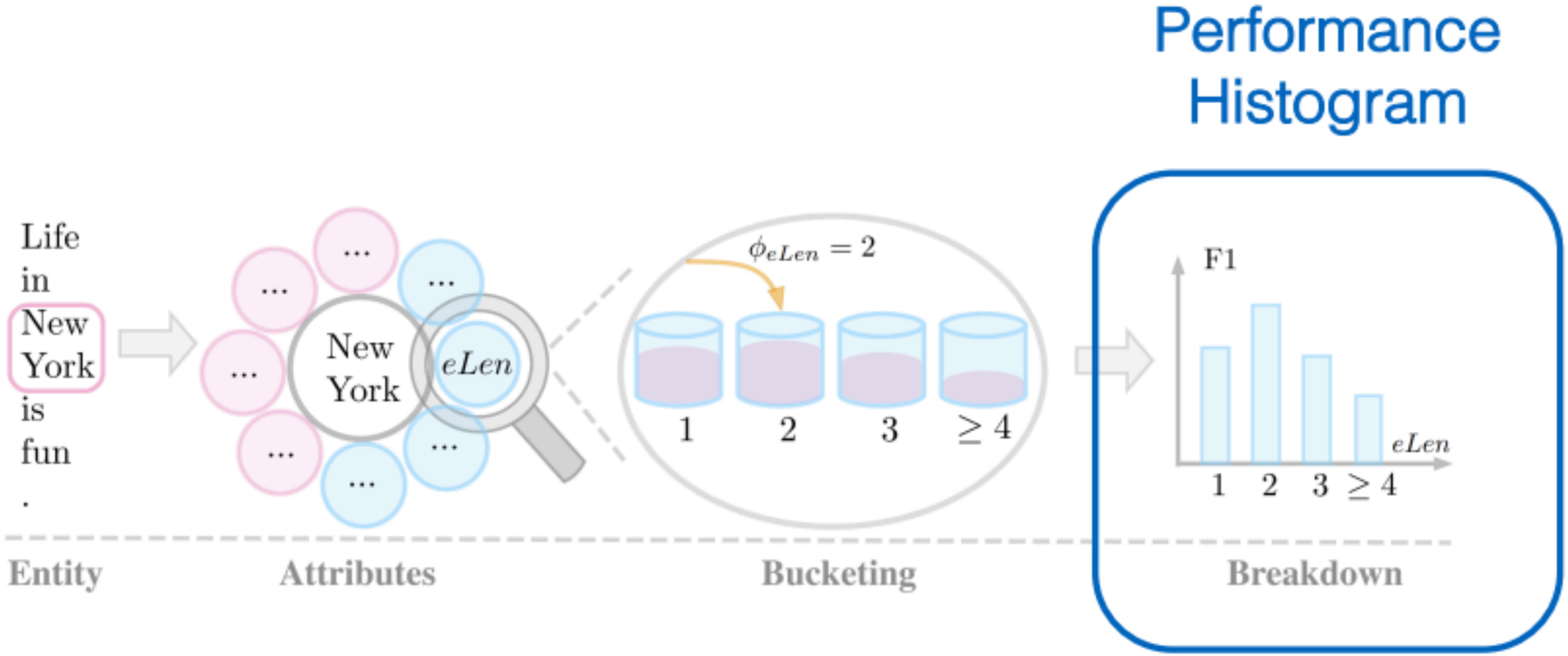
Adapted from slide by Graham Neubig

Defining attributes

- Different tasks could have different attributes
- Token-level, span-level, sentence-level
  - Token-level: part-of-speech tag
  - Span-level: span length
  - Sentence-level: sentence length

Slide credit: Graham Neubig

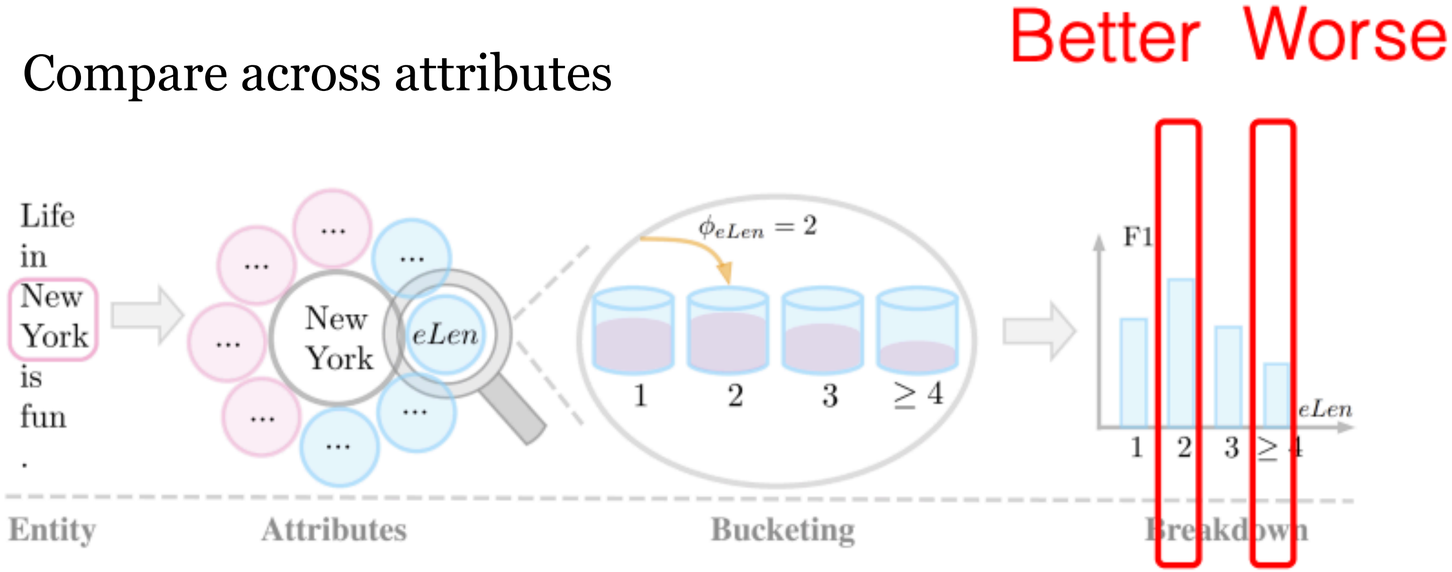
Example: breakdown performance by entity length



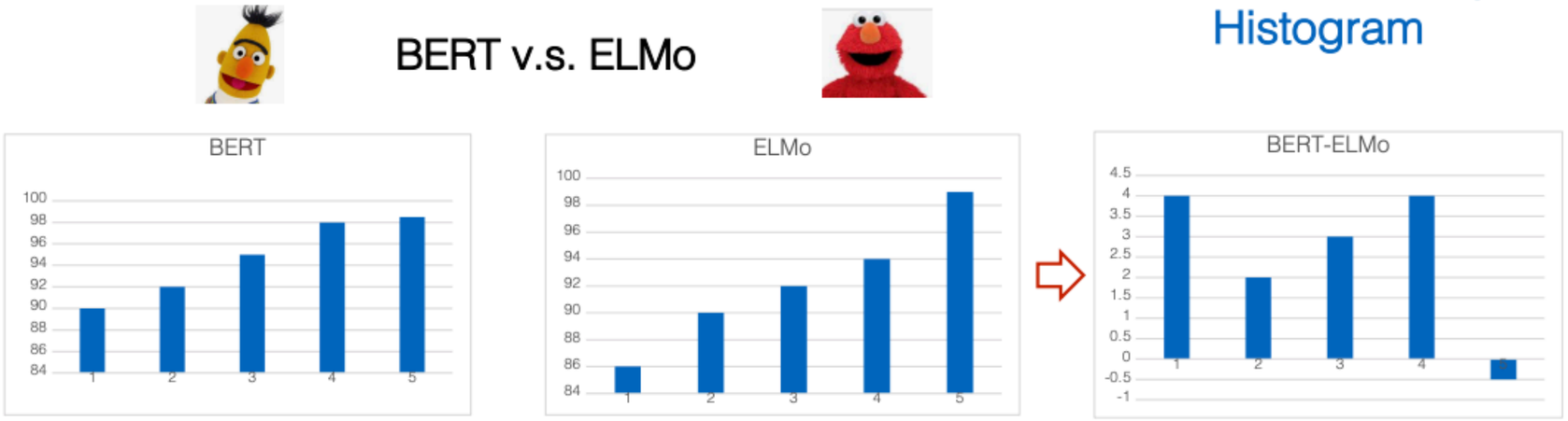
Slide credit: Graham Neubig

Performance histogram

- Compare across attributes



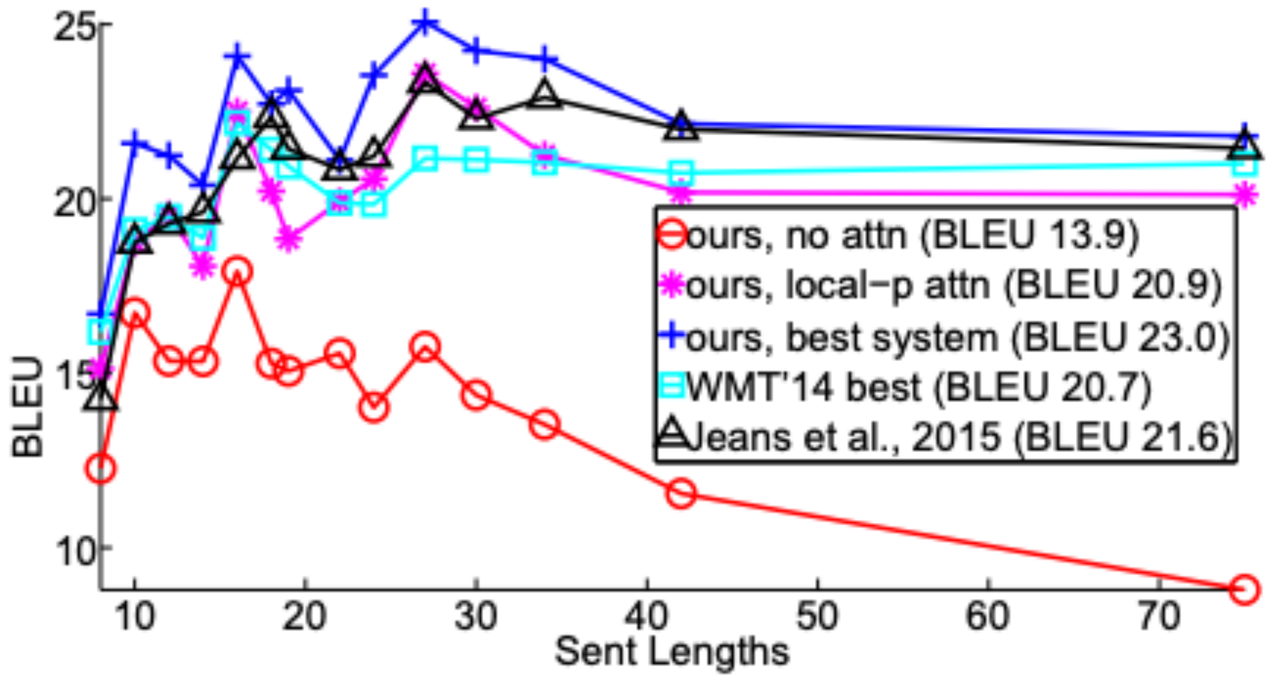
- Compare across models



Slide credit: Graham Neubig

Comparison of model performance by sentence length

- Attribute: sentence length



Effective Approaches to Attention-based Neural Machine Translation
https://arxiv.org/pdf/1508.04025.pdf (Luong et al, 2015)

Integrating unit tests

- Small careful test sets sound like… unit test suites, but for neural networks!
- *Minimum functionality tests:* small test sets that target a specific behavior.

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT***	Labels: negative, positive, neutral		
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
…			
		Failure rate = 76.4%	

- Ribeiro et al., 2020 showed **ML engineers working on a sentiment analysis product** an interface with categories of linguistic capabilities and types of tests.
  - The engineers found a bunch of bugs (categories of high error) through this method!

Beyond Accuracy: Behavioral Testing of NLP models with CheckList
https://arxiv.org/pdf/2005.04118.pdf (Ribeiro et al, ACL 2020)

Slide credit: John Hewitt

Comparisons

Guidelines for making comparisons

- Make sure you have baselines
  - Random
  - Majority Class
  - Conditioned Majority Class
- Other baselines
  - Simpler version of the model vs more complex
    - One layer FFNN
    - Unidirectional/Single-layer RNN
  - Retrieval baseline for generation problems

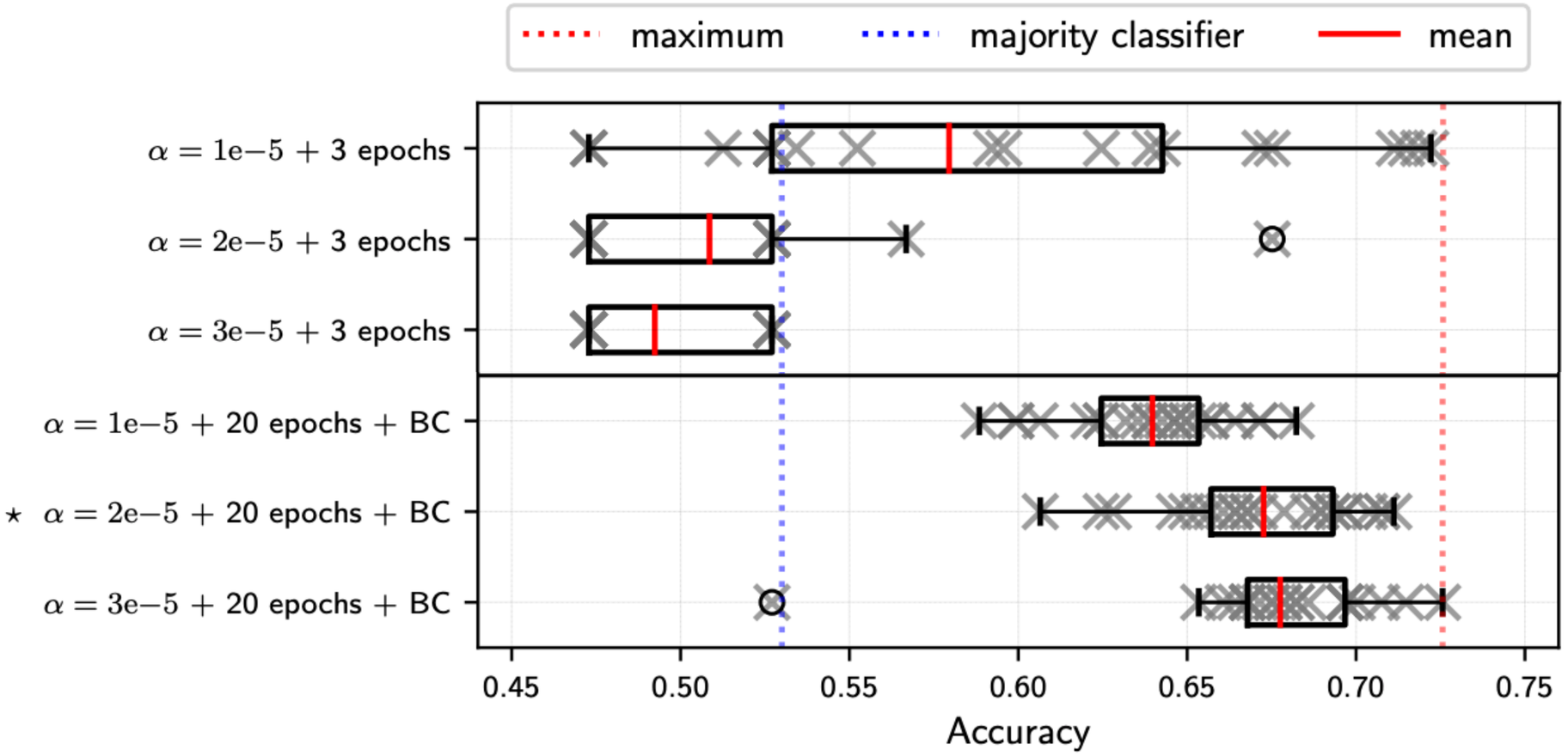
- Always consider: is the complexity warranted?
  - Different types of complexity
    - Simplicity in implementation vs simplicity in concept vs simplicity in model (number of parameters)
  - NN vs rule based vs NBs vs logistic regression

Classifier comparison

- Suppose you've assessed two classifier models. Their performance is probably different to some degree. What can be done to establish whether these models are different in any meaningful sense?
  - Practical differences
  - Confidence intervals
  - Wilcoxon signed-rank test (covered in the sentiment unit)
  - McNemar's test (covered in the sentiment unit)

Slide credit: Chris Potts

Multiple runs with different seeds



On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines
https://arxiv.org/pdf/2006.04884v3.pdf (Mosbach et al, ICLR 2021)

Tips for training and debugging

- TA Tutorial
- Stanford CS231n:
  - https://cs231n.github.io/neural-networks-2/
  - https://cs231n.github.io/neural-networks-3/
- CMU nn4nlp: Graham Neubig
  - http://www.phontron.com/class/nn4nlp2021/schedule/debugging.html
  - https://www.youtube.com/watch?v=KRQHdwpfj-4

Good luck on your project!