Natural Logic Group Meeting

This is a group formed by Prof. Mihai Surdeanu.
Group members: Prof. Mihai Surdeanu, Prof. Eduardo Blanco, Robert Vacareanu and Haris Riaz.

***All meeting notes written by Sushma Akoju.**

5/2/2023

Status and Analysis from Sushma:

- Here are all the results so far (with the default learning rate, default loss but over num_epochs = [4,8] and batch_sizes = [8,16,32]) that were run for 15 folds (since we have 15 SICK examples modified so each folds slides through each of the examples to train, evaluate & test for).
- three tabs in this excel sheet results: qtype refers to Results grouped by Quantifier type, Results grouped by SVO (Subj-verb-Obj), and overall - general results without any grouping for all 15 folds. https://docs.google.com/spreadsheets/d/13OzHa7YwmVkvM2pF25j5aUA8yhwHbtG26KCp98CYDcA/edit?usp=sharing
- There seems to be overfitting which seems too good to be true. But this is inconclusive and leaves room for verifying learning rate and Cross Entropy loss.
- Training for batch sizes 16 & 32 for num_epochs =8 took 4 hours each i.e. 8 hours for Roberta and 8 hours for Deberta just for epochs = 8, batch_sizes = [16, 32]. On the whole for all of the settings, it took nearly 26 hours and while validating and fixing it and rerunning these scripts for nearly ~7 times and fixing bugs along the way to find a stable configuration as well as stable script that really can fine-tune using HuggingFace or PyTorch. **Total efforts towards the fine tuning: 7 * 26 = 182 hours since 12 days is ~15 hours/day.**
- PyTorch direction failed due to reaching the limit of the GPU allotted to ColabPro subscription as it caches a lot of state within GPU and uses HighRAM which is very restricted for ColabPro.
- This is the Google Colab for NLI models chosen: deberta and roberta. The NLI models need to be trained for sentence_transformers models and the script (which is a long but more consistent, stable route to test over this dataset) to verify the results correctly. This is just my opinion and I don't feel confident to say fine-tuning works or not as these experiments seem insufficient. https://colab.research.google.com/drive/15UhNoWI-1cl_QRP9oYriJei4mBOxjrE5?usp=sharing
- for fine-tuned models under the aforementioned settings, it seems insufficient to discuss much about.
- A common, but inconvenient observation: we do not have any stable script despite the fact we have several well-known NLP/Deep learning libraries. So we cannot rely on if the

scripts we finally “got working” are reliable and replicable. Does this seem like a general cause of concern for any NLI model fine-tuning scripts, maybe? - is just my opinion.
- TBD - with Mihai.

Status from feedback and discussion with Mihai:

- [ ] **Mihai:** use lime, check explainability and add more explainability analysis
- [ ] **Sushma:** should we provide entailment, contradiction adn neutral data distribution for imbalanced/balanced accounts? Mihai: we can decide after the deadline.
- [ ] **Mihai:** use 5 folds. entire dataset. -> see if we can finish it.
- [ ] **Learning curve:** 5 folds. show learning curve. with 1 fold it is not learning or not?
- [ ] **Mihai**: print out train+eval+test intersection if there is too much overlap or not -> 1304. This is fixed since we are re-training same model. So instead discard model after each fold test evaluation.
    - [ ] This is because: 1304 -> 1130 (train+eval) and 87 for test. 0 to 1304
    - [ ] fold 1: test: 0:87 -> model trained on 87:1304 including evaluation/validation. And then fold 2: 87:87+87 -> model 87+87:1304 + 0:87
    - [ ] F1-score is 99% -> is too good to be true, certainly overfitting from testing over examples that it was already trained on for exactly same example due to overlap between sliding window vs train sets.
    - [ ] discard the model after each fold -> redo the script .
- [x] ~~**Re-run: experiments for 5 folds and 15 folds.**~~
- [ ] Results for 5 fold Cross validation: https://docs.google.com/spreadsheets/d/1-ElwD8C8rJDY7xOJ5M7DiO-i1qrSL724/edit?usp=sharing&ouid=109002193141570811635&rtpof=true&sd=true

4/15/2023:

- Zero shot results were completed for quantifier type based and SUB-VERB-OBJ SVO based analysis for f1, precision, recall and accuracy. Negation seems to have poor performance of all quantifier types, from sentence transformer models: Deberta and Roberta.

3/23/2023

- [x] ~~Initial analysis - shows deberta and roberta seem to perform better.~~
- [x] ~~More error analysis & update report.~~
- [x] ~~Then fine tune.~~
- [x] ~~Then analyze.~~

3/22/2023

- [ ] Initial analysis - shows deberta and roberta seem to perform better.
- [ ] More error analysis.

- [ ] Then fine tune.
- [ ] Then analyze.
- [x] ~~Update SICK data compositionality data~~
    a. Add monotonicity (upward, downward)
    b. Add Polarity (affirmation i.e. positive polarity, negation i.e. negative polarity)

1/24/2023

1. Mihai to review the disagreements.
2. Generate scores into tables based on categories as suggested by Mihai for zero-shot.
3. Train NLI systems (suggested by Mihai) and verify results and repeat task 2. - validate if the model really learns.
4. Train NLI systems (suggested by Mihai) with changes as per dataset change to teach quantifiers and verify results and repeat task 2. - validate if the model really learns.
5. While step 1 is in progress, work on example rounds 1 & 2 for 2,3,4 tasks.
6. To put the results from 2,3,4 in an Overleaf/Latex document with analysis.

1/10/2023

- Cloud recording of disagreement resolution for Group 2 (Sushma, Robert & Haris) https://arizona.zoom.us/rec/share/UCM5c2Dz7HWGvMXUSsTVPOe_GJm0Onl5JIZ6FIW5zkYAuvCI1YP30dcKv6Nd-Hkw.wYaT3DUzoJX2-WJO

1/9/2023

1. Disagreement_resolution_annotators_group_1 **:** https://docs.google.com/spreadsheets/d/1H1fZgc_m6ClNi-Fd4kK6kN3SsmfYUKrp/edit?usp=sharing&ouid=109002193141570811635&rtpof=true&sd=true
    a. If we cannot comment anything about Intersection, we can choose Neutral. (it could either be empty or non-empty or so)
    b. Also depends on the Universal set chosen for annotations.
    c. Example - Premise: exactly two of dogs are playing by a tree and Hypothesis: two dogs are sleeping by a tree
    Intersection = empty (if they are sleeping, they are not playing).
    Universe = two dogs playing or sleeping by a tree. (minimum set that covers premise and hypothesis).
    Union = every dog
    Selected label = Alternation since Intersection is empty but Union = Universal set
2. Disagreement_resolution_annotators_group_2 : https://docs.google.com/spreadsheets/d/1gsnPHZxF_i9YqAYj6gIXQxjjBn9lTGHT/edit?usp=sharing&ouid=109002193141570811635&rtpof=true&sd=true
3. **Dec 30th Notes from NatLog meeting notes:** two dogs playing by a plant should have been ∃ plant such that plant ∈ plants. Then ∃ plant such that tree ∈ Trees where trees

⊂ plants. two dogs playing by a plant ⊃ exactly two dogs playing by a tree. AS per NLI dissertation chapters 6,

4. **AS per NLI dissertation chapters 6,** *"...The cornerstone of our account is the principle of compositionality, also known as Frege's Principle, after its best-known exponent. In its original form, this principle states that the meaning of a compound expression is a function of the meanings of its parts. In this chapter, we adopt an analogous principle which states that (some of) the entailments of a compound expression are a function of the entailments of its parts. 2 Of course, the principle of compositionality has its critics (notably Chomsky (1975))...."*
5. **Premise:** exactly two of dogs are playing by a tree, **hypothesis:** two dogs are playing by a plant
    a. So we consider H1 = two dogs and P1 = exactly two dogs = two dogs.
    b. P1 U H1 = H1.
    c. P1 Intersection H1 = P1 since "exactly two" means two and only two, P1 should have precendence for intersection in this specific quantifier case, if I understood the parts of premise and hypothesis correctly in our example. then we consider H2 = plant and P2 = tree.
    d. P2 U H2 = H2 since tree ⊂ plant i.e. P2 ⊂ H2.
    e. P2 intersection H2 = P2 (tree).
    f. Universal set = two dogs playing by plants.
    g. P1 U H1 + P2 U H2 = H1+H2. P1 Intersection H1 and P2 intersection H2 = P1 + P2. but we have only "a plant". since intersection is nonempty and union is not a universal set. Hence I assumed this has to be neutral but I am really not sure of my annotation for this example.
6. I have one more question.
    a. premise: a boy is standing in the cold water, hypothesis: every boy is standing in the water.
    b. P Intersection H = boy standing in cold water.
    c. P U H = All boys standing in water (warm, hot, cold, and ice water) = Universal set.
    d. looks like Cover.
7.

1/8/2023

- ☐ **NLI System analysis (Sushma): WIP to be shared by tomorrow 10am.**
- ☐ **To fine tune a best NLI system over modified sentences. (using cross validation) : ON HOLD**
- ☐ **Important Note about Ground Truth labels:** Since the inter-annotator agreement analysis revealed high disagreement, we paused the NLI system analysis. This is because we need Ground Truth labels to finetune or train an NLI system. But we do not have "Ground Truth" labels for complete annotations for full, modified SICK examples dataset.

- [ ] For Group 2, which was primarily annotated by me, Robert and Haris: I created this excel sheet with rows with disagreements highlighted in Yellow so me, Robert and Haris could provide justification for the interpretation and decision for the label. https://docs.google.com/spreadsheets/d/1gsnPHZxF_i9YqAYj6gIXQxjjBn9lTGHT/edit?usp=sharing&ouid=109002193141570811635&rtpof=true&sd=true
    - [ ] We informally planned to meet on Tuesday 11am Tucson time (among various intersection of times suggested)
- [ ] For Group 1's disagreements, we meet on Monday the Jan 9th @ 9am to discuss the disagreements.
- [ ]

1/4/2023

- Full Inter-annotator agreement analysis: https://github.com/sushmaakoju/natural-logic/tree/main/notebooks/combined_annotations
- For each individual statistical explanation and analysis: the notebooks for group 1 and group 2 annotations are:
    - Group 1: https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/combined_annotations/group1_inter_annotator_agreement_analysis.ipynb
    - Group 2: https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/combined_annotations/group2_inter_annotator_agreement_analysis.ipynb
- Since for each group, the inter-annotator agreement was very low, i.e. less than 40% unfortunately.
- WE enforced a new "Disagreement resolution" to record justification for each of the disagreements and see what comes out of this analysis, as per suggestion from MIhai.
- For Group 2, which was primarily annotated by me, Robert and Haris: I created this excel sheet with rows with disagreements highlighted in Yellow so me, Robert and Haris could provide justification for the interpretation and decision for the label. https://docs.google.com/spreadsheets/d/1gsnPHZxF_i9YqAYj6gIXQxjjBn9lTGHT/edit?usp=sharing&ouid=109002193141570811635&rtpof=true&sd=true
    - We informally planned to meet on Tuesday 11am Tucson time (among various intersection of times suggested)
- For Group 1's disagreements, we meet on Monday the Jan 9th @ 9am to discuss the disagreements.

12/28/2022

TODOs: Annotations & Programming work:

- [ ] **Observations during annotations:**
  - [x] ~~**Mihai:** About Contradictions - after modification, can be trickier.~~
  - [x] ~~**Sushma** - I corrected the verb modification for the following sentence (bug- adverb +verb instead of verb + adverb and modifier subject instead of part of subject of leftmost nominal modifier in Prepositional phrase).~~
    - [x] ~~SICK_ID 342: Premise: a girl with a black bag is on a crowded train. Hypothesis: a cramped black train is on the bag of a girl - needs corrections.~~
    - [ ] Look for automating "grammatical" correctness
- [ ] **Corrections, checking script reproducibility: Mihai's guidance:** The corrections should be automated by camera-ready time for the paper. During submission time, it's Ok to correct some data manually for this dataset paper.
- [x] ~~Ground Truth by voted majority (suggested by Mihai) - for annotations WIP~~
- [ ] **NLI systems - best performance - last step**
- [ ] **None of the scripts were reviewed by Haris or Robert yet. TBD.**
- [ ] **NLI System analysis (Sushma): WIP to be shared by tomorrow 10am.**
- [ ] **To fine tune a best NLI system over modified sentences. (using cross validation) : WIP**
- [ ] **Possible direction to explore: to collect scores**

To note the Status for the paper to be written towards this:

- [ ] Create a template to add:
  - [ ] Annotations scheme
  - [ ] Inter annotator analysis
  - [ ] Ground truth labels & analysis
  - [ ] NLI system performance analysis
  - [ ] Fine tune best performing model (depends on NLI system performance)
  - [ ] Data section
  - [ ] Add Assumptions
  - [ ] Abstract
  - [ ] Motivation
  - [ ] Chat-GPT2 - Eduardo
- [ ] Template is for short paper @ ACL
- [ ] First Deadline target: Jan 13th 2023 (ACL Abstract submission)

To note the code written towards the paper so far "Completed":

1. **Initial noun phrase modification (Robert + Sushma)**: Robert guided and mentored thow to use Berkeley Neural Parse and get the leftmost Noun phrase from the dependency parses. (first part of this script: sentence_modifiers_annotations.ipynb contains a block on subject/noun phrase modification. *line 18: get_leftmost_np(sentence)*)
2. **For the verb phrase modification: (Sushma)** I implemented it by myself using Berkeley neural parse in python. (later part of this script: sentence_modifiers_annotations.ipynb contains block on subject/noun phrase modification)
3. **For Object modification: (Sushma)** I used Scala + Processors api to extract specific patterns.
   a. For this task, I analyzed and found that there are different types of rightmost NPrase POS tag sequences: "DT+NN", "DT+NNS", "JJ+NN", "JJ+NNS", "DT+JJ+NN", "DT+JJ+NNS", "DT+NN+NN", "IN+NNS", "VBZ+JJ", "IN+DT+NN", "IN+DT+NNS".
   b. So each of the specific cases is handled in this case. We also observe a Prepositional Phrase case which is a rather complex one.
   c. Also attempted to generalize the "modifier" process, however patterns cannot be generalized, depending on the type of sentences.

4. **Selecting adjective modifiers for Noun Phrase modification:** discussion with Robert. But we did not select the modifiers that were discussed from this.
5. **NatLog meeting approved Modifiers:**
    a. **Noun phrase modifiers:** 'every', 'some', 'at least','not every','exactly one', 'all but one','everyone of', 'no' and adjectives: "green", "happy", "sad", "good", "bad"
    b. **Verb phrase modifiers:** “not”, adverbs: "abnormally", "elegantly", "always", "never",
6. **First round of inter annotator agreement analysis (Sushma):** Round 1: inter_annotator_agreement_analysis.ipynb
7. **Second round of inter annotator agreement analysis (Sushma):** Round 2: inter_annotator_agreement_analysis_round2.ipynb
8. **Sentence, Modifier, Countwise binning:** histogram_counts.ipynb
9. **Auto generating annotator blocks for each sentence modified from SICK dataset: (Sushma)** https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/auto_generate_annotator_files.ipynb
10.

12/27/2022

- Corrected following sentence premise and hypothesis for verb modifiers: (bug adverb +verb instead of verb + adverb as well as and modifier subject instead of part of subject of leftmost nmod before cop) :
- SICK_ID 342:
- Premise : a girl with a black bag is on a crowded train, Hypothesis:a cramped black train is on the bag of a girl
- **Bug:** a girl with a black bag **not is** on a crowded train
- **After manual correction:** a girl with a black bag **is not** on a crowded train

12/24/2022

- Feedback from Dec 23rd from Mihai: uploaded changes https://github.com/sushmaakoju/natural-logic/tree/main/natlog
    - ☑ ~~“t h e” bug~~
    - ☑ ~~do not replace JJs, append instead: “the good cold water”~~
    - ☑ ~~if prepositional phrase, modify the left-most NP~~
        - ☑ ~~for the above, add the pattern: IN+DT+NN~~
- Observation: while I was running an “analysis” of SICK dataset and reading the SICK paper: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf it turns out we would not have strict FE,RE pairs of premise, hypothesis in SICK unless sentences are longtail.
- Interesting aspect: SICK paper discusses Meaning Preserving Transformations, which are essentially rules and/or with POS pattern sequences. Reference section: 3.1. Sentence Normalization.
- So this led me into following:
  SICK dataset already accounts for FE and RE for each pair of sentences i.e. given, A, B sentences - entailment A -> B as well as B -> A was labelled.
  Secondly, for example in our selected example: we have
    - **premise= an old man is sitting in a field**,
    - **hypothesis = a man is sitting in a field**
    - the column values

        - entailment_AB : A_entails_B and implies this is FE
        - entailment_BA : B_neutral_A - this implies Neutral other way around
        - SICK label: Entailment
        - We planned to flip and add FE examples to look for RE but it is not the case unless sentences have to be long and were “expanded” as per their discussion in the paper.
        - So now we have only 15 examples.
- Script to auto generate annotator files based on sliding window with stride (overlap) https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/auto_generate_annotator_files.ipynb
- I created the annotator sentences and wrote script to generate by using an approach suggested by Mihai:
    - ☑ ~~Each of the 5 annotators annotate 8 example sets with an overlap of 3 to 4 examples~~
    - ☑ ~~What we mean by example set is original premise, hypothesis pair + all of modified examples just for this one example.~~
    - ☑ ~~We name each SICK example out of 15 examples as block numbers.~~
    - ☑ ~~So each annotator file has 3 to 4 example overlaps but each of the 8 examples is a separate sheet in the same excel sheet.~~
    - ☑ ~~Annotator 1 excel file -> 8 sheets named as block_{num}_{sick_id}~~

12/22/2022

1. Quick updates:
   Upcoming tasks for everyone: we have 1770 sentences to annotate. Original sentences: 20. So we each have 8 sentences and their corresponding modified sentences.
2. For all work related to implementation:
   https://github.com/sushmaakoju/natural-logic/tree/main/notebooks
3. Our NLI_XY++ data generated:
   https://docs.google.com/spreadsheets/d/19-XIdK32aGsToPCXOhWEQ8Q0IuBfc8KR/edit#gid=304679586

About premise, Hypothesis pairs - counts:

SICK # : 20	**Count**	**# modified**	**Premise, hypothesis modifications**	**Final #**
Object modifiers	**12**	**12 * 20 = 240**	**3 * 240 = 720**	**670**

Subject Modifiers (NP)	**13**	**13 * 20 = 260**	**3 * 260 = 780**	**780**
Verb Modifiers (VP)	**5**	**5 * 20 = 100**	**3 * 100 = 300**	**300**
Original	**20**	**-**	**-**	**20**
Column #	**50**	**600**	**1800**	**1770**

Should have been : 1820 but after filtering out some: we have 1770.

12/19/2022

- [x] ~~Strategy: Append Not replace adjective for the POS Tag sequence -> "JJ +NN"~~
- [x] ~~VBZJJ = Value("VBZ+JJ")~~
- [ ] Include Chat GPT2 only - not GPT2
- [ ] Attempt to complete analysis by early new year
- [ ] Create a structural Template for ACL paper referring: Chapter 1 guidelines from https://github.com/clulab/clulab/wiki/Structure-of-Academic-Papers-&--Common-Writing-Issues
- [ ] Dataset paper Guidelines for inspiration: https://arxiv.org/pdf/2211.16492.pdf
    - [x] ~~Refer analysis sections~~
    - [ ]
- [x] ~~Early next week for sending all modified sentences for annotations - goal~~
    - [x] ~~To check all sentences once again~~
    - [x] ~~So far full dataset of modified sentences are here: https://docs.google.com/spreadsheets/d/19-XldK32aGsToPCXOhWEQ8Q0luBfc8KR/edit#gid=304679586~~

- [ ] **Sushma** completed the rest of the tasks from 12/1 (were on a slower pace due to final exams, projects and final homeworks.). - 80% completed
- [x] ~~Object modification rules for SUBJ+VERB+OBJ => https://github.com/sushmaakoju/natural-logic/blob/main/natlog/src/main/scala/ObjTagSequence.scala~~

Low priority:

- [ ] In Jan 2022 - to resume after ACL deadline -> **ProofVer - NatLog**
- [ ] **NL to FOL - third, but parallel priority**

12/1/2022

Meeting with Mihai, Eduardo, Robert, Haris and Sushma:

- [ ] We follow Bill MacCartney's NLI dissertation as a standard
- [x] ~~Round1 vs round 2 agreement~~

    a. https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/inter_annotator_agreement_analysis.ipynb
    b. https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/inter_annotator_agreement_analysis_round2.ipynb
- ☑ ~~Add 3 more columns~~
    a. "modifier (every) ,
    b. What is being modified? premise or hypothesis or both
    c. Which part of the premise/hypothesis is being modified?
    d. https://docs.google.com/spreadsheets/d/19-XIdK32aGsToPCXOhWEQ8Q0IuBfc8KR/edit#gid=304679586
- ☑ ~~Generate histogram by modifier, histogram by premise modifier, by subject, by object~~
    a. https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/histogram_counts.ipynb
- ☑ ~~Round 3 annotations - just modify noun phrases~~
    a. Example: boy is hitting a every baseball
    b. Skip the verbs (just modifying objects)
    c. For every premise/hypothesis we modify object i.e. baseball and we just consider
    d. P: a boy is hitting a baseball, H: a child is hitting a baseball
    e. https://docs.google.com/spreadsheets/d/1EO6XHlbnDAqt-96RZ6glCm6fszEuH9q0RGW-FzarRcY/edit?usp=sharing
- ☑ ~~Full dataset annotations Plan~~
    a. We can include baseline in final dataset (~~20~~ 15 examples selected from SICK dataset)
    b. Check all sentences labels from SICK dataset
    c. Split data - assign 8 to each person, each set of 6 assigned to 2 people
        - ☑ ~~5 people - 2 sets of 4 overlapping between each pair of annotators~~
        - ☐ 1769 sentences=> 1769
- ☐ NLI systems - performance
- ☐ Switch Premise and Hypothesis and get two-way score [6]
    a. topmost of 6
    b. Consider “Entailment, Contradiction, Neutral”

    Prediction Conflict resolution strategy
    c. FE - Entailment both ways
    d. Contradiction - Negation union Neutral
    e. But we get FE, RE, Neg, Neutral
        - ☐ Both directions [6 score, highest]
        - ☐ 50%, 40% and 10% - select just 50% (topmost)

- Later stage - consider ensemble strategy - for discussion for later

11/23/2022

- ****** Thanksgiving *****

11/17/2022

Meeting with Mihai, Eduardo, Haris and Sushma:

- Whiteboard notes/guidelines for next round of annotations: https://arizona.box.com/s/jb8pznl8oevjsahkinu228h6tf8cax2t
- The guidelines are updated here: reviewed from Mihai https://arizona.box.com/s/azghlefy2maoujx1ystccpqf68m5ogbf
- Second round annotations: https://arizona.box.com/s/ad4t86261vd2z9470j03l2usulw8u182
- Mihai to review the guidelines document per this morning's discussion, review the inter-annotator agreement and next round example annotations guidance.
- Additionally, we compared annotations between Mihai and Eduardo's labels.
- We discussed Robert's comments.

About Inter annotator agreement analysis:

- Total 35 * 5 = 175 labels required by 5 annotators for 35 premise-hypothesis pairs. If Cover, Alternation and Equivalence cases were marked to be Neutral, then Neutral counts change to 47+6 = 53.
    a. Forward Entailment (FE) - **105**
    b. Reverse Entailment (RE) - **19**
    c. Neutral - **47**
    d. Negation - 6
    e. Cover - 3
    f. Alternation - 0
    g. Equivalence - 3
- **Cohen's kappa ~50.3%**.
- **Scott's pi:** 50.29% (would be the same as Cohen's Kappa Score).
- **Tasks for Sushma:**
    - ☑ ~~Complete inter annotator analysis and share with Mihai~~
    - ☑ ~~Upload to Github: https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/inter_annotator_agreement_analysis.ipynb~~
    - ☑ ~~Email Mihai with status on follow up tasks~~
    - ☐ Send selected SNLI models list for next analysis
    - ☑ ~~Send examples to Mihai for another single example with 38 modifiers set for second round of annotations based on new guidelines~~
- Updated documents and excel sheets as per today's meeting guidelines:
    a. NLI annotation task guidelines - updated, to be reviewed by Mihai - Box Note
    b. Combining first round of annotations for an example : excel sheet
    c. excel sheet: reference tabs modified-verb-phrases and modified-noun-phrases for selecting example set for next round of annotations
    d. https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/inter_annotator_agreement_analysis.ipynb

11/16/2022

EB comments for tomorrow:

- I cannot tell if mentioning to “A man” in the P and H refer to the same man or not.
    - “A man wearing a red jacket” is neutral with “A man wearing a black jacket” unless we assume that both are describing the same situation in which there is only one man. SNLI makes this assumption but I don’t know if we should.
- Same than above but with “Some men” in the P and H. Is that referring to different kinds groups of men?
    - “Some men sitting in the ground” entails that “There are men not sitting in the ground”? I am not so sure but we should be consistent.
- Negation is tricky. I don’t know if we are looking for “real” entailments “or reasonable inferences / implicatures”.
    - “John doesn’t drive a fast car” to me “(sort of) entails” that “John owns a non-fast car” (despite it is not necessarily the case: you can drive a company car and “not driving a fast car” does not really entail “driving a non-fast car”.

11/15/2022

1. Noun phrases modifiers: 'every', 'some', 'at least','not every','exactly one', 'all but one','everyone of', 'no' and adjectives: "green", "happy", "sad", "good", "bad"
2. Verb phrase modifiers/adverbs: "abnormally","elegantly","always","never", “not”
3. Excel sheet for annotations (for Inter annotator agreements): https://arizona.box.com/s/w2fxwg6i6k9evuw9tradaf6kk8v0mgod
4. Annotations email

11/14/2022 - 11/15/2022

- ☐ Modifying verb Phrases using processors - **11/13/2022 & 11/14/2022** **IN PROGRESS**
- ☑ ~~every one -> everyone **11/15/2022**~~
- ☑ ~~Sushma: to check for syntactic correctness using perplexity score : **11/15/2022**~~
- ☑ ~~**Meeting with Mihai - completed**~~
- ☐ Colab Notebook for this task: https://github.com/sushmaakoju/natural-logic/blob/main/notebooks/sentence_modifiers_annotations.ipynb
- ☑ ~~To send an email to Mihai and everyone about annotation task for inter annotator agreement for labelling NLI as per Bill MacCartney’s NLI dissertation.~~
- ☑ ~~Email everyone with excel sheet for annotations: https://arizona.box.com/s/w2fxwg6i6k9evuw9tradaf6kk8v0mgod~~

11/11/2022 - 11/13/2022

- [x] ~~Modify sentence by only adding adverb before the verb **11/15/2022**~~
- [x] ~~9 generalized quantifiers : 'every', 'some', 'at least','not every','exactly one', 'all but one','every one of', 'no' - **11/13/2022**~~
- [x] ~~9 adjectives: "green", "happy", "sad", "good", "bad" **11/13/2022**~~
- [x] ~~Grammatical correctness: add “a” or determiner if it is not already there~~
- [x] ~~Corrections for verb phrases using Spacy / Neural Parser -**11/14/2022**~~
- [ ] Modifying verb Phrases using processors - **11/13/2022 & 11/14/2022** **IN PROGRESS**
- [x] ~~every one -> everyone **11/15/2022**~~
- [x] ~~Sushma: to check for syntactic correctness using perplexity score : **11/15/2022**~~
- [ ] To send an email to Mihai about following for going ahead and sharing with rest of us:

11/11/2022

Meeting with Mihai, Robert, Haris and Sushma:
Tasks for Sushma to be completed by end of 11/11:

- Modify only single position
- 9 generalized quantifiers : 'every', 'some', 'at least','not every','exactly one', 'all but one','every one of', 'no'
- 9 adjectives: "green", "happy", "sad", "good", "bad"
- Grammatically correctness: add “a” or determiner if it is not already there
- every one -> everyone
- Sushma: to check for syntactic correctness
- To send an email to Mihai about following for going ahead and sharing with rest of us:

  Annotation tasks:
  - 4 examples each from 20
  - Don't look at each other’s work
  - Labels produced will be larger
  - Approach
  - Pairwise inter annotator agreement: Kappa, Fleiss -> Sushma
  - Labeling approach from SICK labels to NLI labels: FE, RE, NEG, Contradiction, (Independence/neutral, Cover/Alternation)

  NLI Models:
  - AllenNLP Roberta Large
  - NLI models from HuggingFace
  - Uncased vs cased
  - Just add the text (no spaces, or whitespaces at beginning or ending)
  - Use pipeline

    - By wednesday
    - Do we need probing experiments like NLI_XY or not?

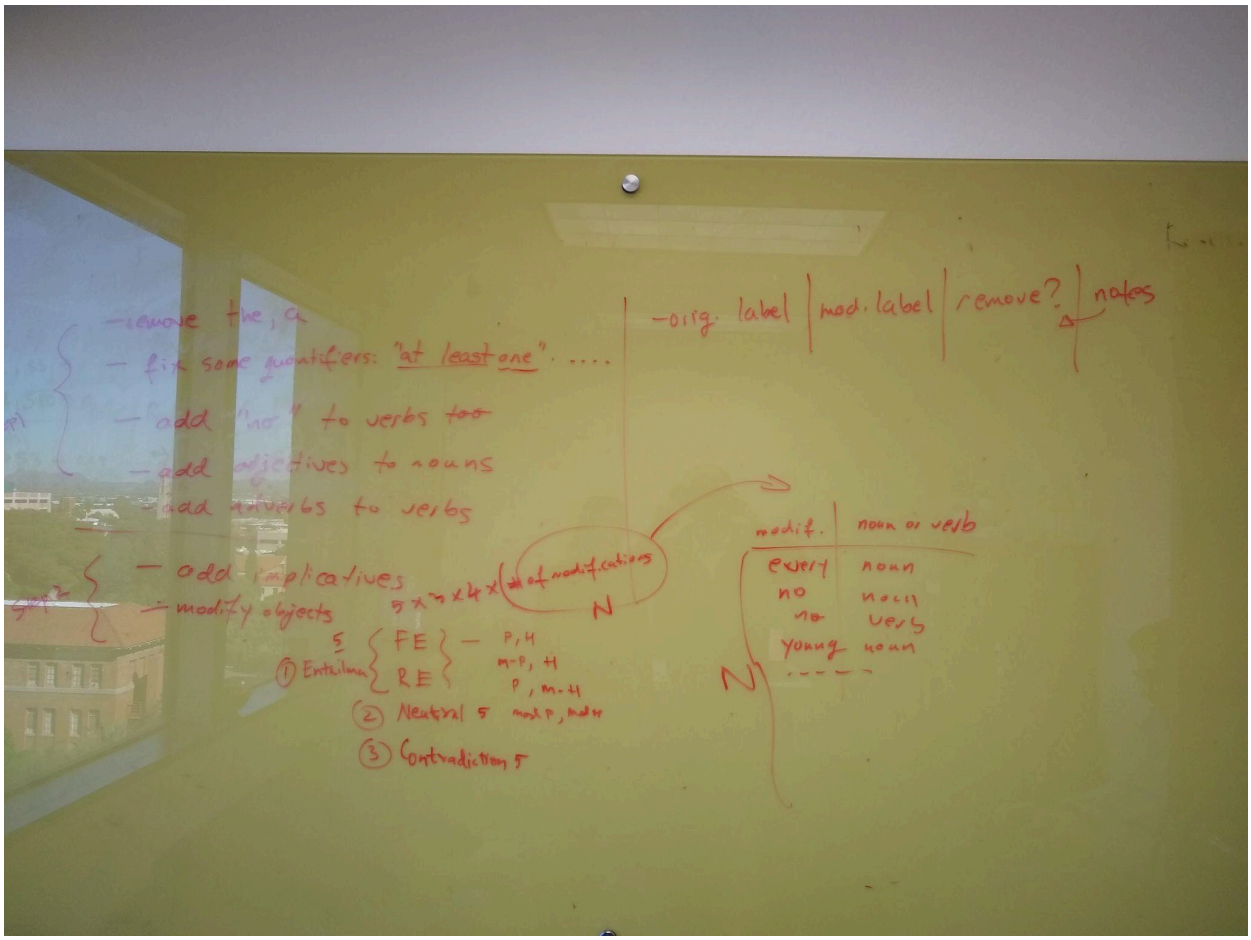
11/4/2022 - 11/09/2022

- Worked on tasks since. Each of the tasks' status was updated.
- 20 sick examples were selected
-
-
- 14 each of premises and 19 hypotheses after modification
- 20 * 13 * 4 = modified premises and 380 modified hypotheses including original one.
- **How to match Premise and Hypothesis?** Out of 380 premises, I matched for each premise, each of the hypotheses. So a total of 380 * 380 are possible candidates. Should we change the matching strategy? Since there are only so many valid premise, hypothesis cases.
    - Modified premise, hypothesis
    - Premise
-
- **Next task: for each modification, for each combination of premise, hypothesis, remove invalid combinations as per projectivity and join signatures.**
- For adjectives, the rule to add modifier is:
    - If there was a Cardinal number (CD) which is first word in sentence,
        - If the modifier is a determiner or any other modifier other than Adjective
            - Add modifier before CD and append "of" after the CD
            - Also remove "one" from modifier if modifier ends with a "one" such as "exactly one"
        - If the modifier is an adjective
            - Add the adjective after the CD

- Integrating **processors** to generate sentences:
    - Initially worked for generating single premise, but
    - Loops seems to have been slightly cumbersome and
    - Probably I chose between a couple of ways to modify and save to a Seq vs using Breeze methods (breeze may not be necessary)
    - **Still in-progress**
    - But generating sentences with modifiers is more easier
    - Plus I could add a rule eventually to just make the right modification? - is to be worked on
- Generated sentences as of today are from Neural parser
    - Applied all of the rules
    - Sentences are placed at: https://docs.google.com/spreadsheets/d/1XHtCy_uLV0BnnMXIpvyTtCPdWteqJNFEqo9phIGOzEk/edit?usp=sharing
- **Adverb selection mechanism:**

    - Abnormally
    - Oddly
    - Elegantly
    - Ferociously
    - Always
    - Never
- **Approach suggested by Robert** for adverb selection process to extract most “relevant” adverbs for the given verbs from sentences
    - LM - to decide adverb, verb combinations
    - By using perplexity score to select most appropriate adverbs
    -

11/4/2022



Meeting with Mihai, Robert, Haris and Sushma:

Tasks for Sushma

- ☑ ~~Step 0: send a table with all of the modifiers and if the modifier holds for noun and/or verb~~

modifier	verb or noun
every	noun
some	noun
at least	noun
not every	noun
exactly one	noun
all but one	noun
every one of	noun
not	verb
no	noun
no	verb

- [ ] Step 1: (we focus on modifying the subject in this step)
  - [x] ~~Remove ***the***, ***a*** modifiers~~
  - [x] ~~Fix some of the modifiers “at least one” for example~~
  - [x] ~~Add not to verbs too~~
  - [x] ~~Add adjectives to nouns~~
  - [ ] Add adverbs to verbs
  - [ ] To select adverbs for the verbs in Verbs: Sitting, Standing, Hanging,Climbing,Hitting, Playing, Throwing, Riding, Jumping
  - [ ] Adverb candidates
    - [ ] Abnormally
    - [ ] Oddly
    - [ ] Elegantly
    - [ ] Ferociously
    - [ ] Always
    - [ ] Never
  - [ ] Approach suggested by Robert to use
    - [ ] LM - to decide adverb, verb combinations
    - [ ] By using perplexity score to select most appropriate adverbs
  - [x] ~~Finally for each of 5 examples from SICK dataset:~~
    - [x] ~~Entailment (we want FE, RE - 5 each)~~

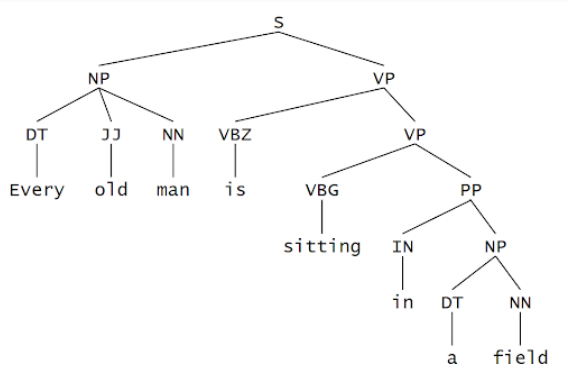
        - [x] ~~Contradiction~~
        - [x] ~~Neutral~~
    - [x] ~~So we have 5 * 4 = 20 examples from SICK dataset~~
    - [x] ~~For each of the 20 examples, create following 3 pairs 20 * 3 = 60:~~
        - [ ] not Premise, Hypothesis
        - [ ] Premise, not Hypothesis
        - [ ] Not Premise, not Hypothesis
    - [x] ~~For each of the 15 examples, consider annotate 4 labels from following, inline with Bill MacCartney's dissertation:~~
        - [x] ~~Forwards Entailment (FE)~~
        - [x] ~~Reverse Entailment (RE)~~
        - [x] ~~Negation (Neg)~~
        - [x] ~~Neutral (Cover)~~
    - [x] ~~For the selected modifiers which total to some N, for each of the 60 examples, add modifiers. 60 * N should be the total number of sentences.~~
    - [ ] Use processors generate sentences - errors
    - [ ] To share the sentences with each one of us before next week's meeting.
    - [ ] To add
- [ ] Step 2:
    - [ ] Add implications
    - [ ] Modify objects
- [ ] Tasks for each one of us:
    - [ ] We would have Robert, Haris, Mihai and Sushma as annotators.
    - [ ] We would have a preliminary set of examples - to assess the inter-annotator agreements.
    - [ ] We follow this approach to label one of the labels: FE, RE, Neg and Neutral
        - [ ] If-then approach (is simple for decision making)
        - [ ] We can follow Projectivity or join signatures and a few more concrete calculus formalizations from Bill McCartney's dissertation for a formal approach to decide the labels.
        - [ ] We can follow the alignment algorithm approach from NLI dissertation
        - [ ] We can also evaluate the label decision for self-retrospection by using the First order logic approach.
        - [ ] We can also add more approaches we discovered during annotations
        - [ ] If any of us decide to remove an example sentence from the dataset, we may

- Evaluation mechanism:
    - We may use probing approach similar to NLI_XY paper
    - We run each of the sentences over Roberta and find out maybe all of the quantifiers are correctly labeled/predicted by Roberta.

- A natural future direction might be just an alignment algorithm if the NLI systems do not label the examples from quantifiers.

11/3/2022

- Sentences generated from following implementation are placed here: https://docs.google.com/spreadsheets/d/1XHtCy_uLV0BnnMXIpvyTtCPdWteqJNFEqo9phlGOzEk/edit?usp=sharing
- Implemented the full but simple auto-generated sentences with modifiers.
- https://colab.research.google.com/drive/1Ks5EQMeIQppIDPNUANgrTLuhNmwqBpaP?usp=sharing



-
- modifiers = "some", "every", "an", "the", "at least", "not every", "exactly one", "all but one"
- and this_premise = 'Every old man is sitting in a field.' In this case we just modify Determiner to one of the modifiers.
- But suppose we do expand the list of modifiers to "Many", "Most", "Few", "Several", atleast we have one rule: then in that case we modify the first NN in NP -> man -> men and also leftmost VBZ in VP i.e. is -> are
- To generalize the above rule to vast number of verbs and respective verb forms, we need the verb from the current premise sentence, we find a tense we need for the selected modifier and also modify the first NN in Noun Phrase to plural form i.e. NN to NNS and NNP to NNPS.
- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- So we just consider: "some", "every", "an", "the", "at least", "not every", "exactly one", "all but one"

More Papers on NL to FOL & logic puzzles:

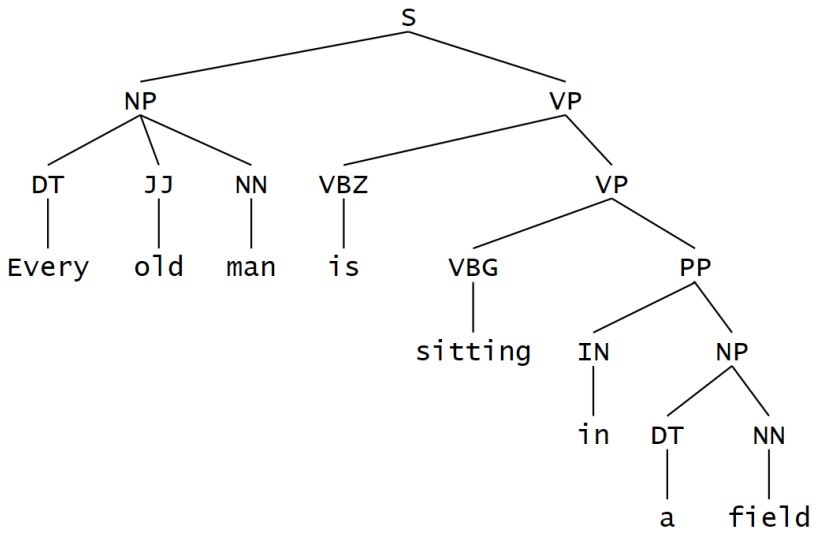
- Existing paper on Neural machine translation "english" to first order logic: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9401852
  - https://github.com/alevkov/text2log and they have got good accuracy
  - They seem to have used a CCG2Lambda parser to generate logical formalisms (similar to Langpro)

    - Quoting this from the paper: "A valid candidate for providing ground truth labels, is a CCG-based formal semantic parser (ccg2lambda) that generates logical formalisms from English sentences and outperforms state-of-the-art first-order systems [21]."
    - The data is split into x & y (x - all english sentences) and y consists of all generated FOL statements from ccg2lambda.
    - The repository for ccg2lambda: https://github.com/mynlp/ccg2lambda
    - To add this paper to the Survey on translation from NL to FOL
- https://aclanthology.org/2020.findings-emnlp.100.pdf Distill-Bert model can solve logical puzzles using quantifiers "comparatively".

11/2/2022

Meeting with Robert

- Rules and discussion as per guidance from Robert:
- For the purpose of discussion, we consider just this one sentence for rules for negation as well as adding modifiers as extracted using Neural Parser from SpACY (https://colab.research.google.com/drive/1Ks5EQMeIQppIDPNUANgrTLuhNmwqBpaP?usp=sharing):



-
- Every old man is sitting in a field
    1. **First look for Noun Phrase (NP) connected to S node**
       Left most NP -> negate DT/change DT
       Then if NP has a DT, modify this (Some, every, No, An, The)
       For Negation case, give DT preference over VP

    2. **Now for the case where Verb Phrase VP,**
       Only negate the VBZ only if DT is the or an

3. We want avoid following “weird sounding” sentences for negation:

   Every old man is NOT sitting in a field
   Some old man is NOT sitting in a field

4. But following seems valid:
   The old man is NOT sitting in a field

   So we only negate VB in VP only if DT in NP is “The” for example. There could be other determiners.
   For example, my friend is playing baseball.

5. We do follow points 3 and 4, since this is a case where we do not want to have an example for a 1 in 1000 case when instead this example should have been 1 in a million data samples. Just to make sense of good sampling.
6. We can discuss if we do prefer to include examples from 3 and 4.
7. The most important reasoning here is that when we “auto-generate” negation over Noun phrases or verb phrases, we do end up creating sentences that may not be used in real despite their syntactic correctness.

- Implemented partially some parts for one example.

10/31/2022

Created this Colab Notebook with Neural parser from spacy:

- https://colab.research.google.com/drive/1Ks5EQMeIQppIDPNUANgrTLuhNmwqBpaP?usp=sharing
- Extracted the sentence examples to csv and for each premise hypothesis sentences, generated parse trees and saved
- worked on pre-order traversal of the parse tree for one example (recursively), since the recursion here is not too long.

10/28/2022

Meeting with Robert about parser:

- Robert explained about the concept of how the heads can be extracted from VP
- Intuition about “**heads**” from Noun phrases and Verb phrases: https://dictionary.cambridge.org/us/grammar/british-grammar/heads
- Robert suggested to use this Neural parser : https://spacy.io/universe/project/self-attentive-parser
-

10/27/2022

Meeting with Mihai, Robert, Haris and Sushma

1. Try in both direction for all reverse entailment
2. Is the modifier a noun or verb ?
3. Sushma's Tasks
    a. 5 of each of Forward, Reverse and Equivalence - from SICK - and share with Mihai
    b. Add 3 negation - 3 cases (Not P & H, P & Not H and Not P & Not H)
    c. Find the first verb phrase and find the heads and apply modifiers to the heads
    d. Is AllenNLP keeping the same labels?
    e. Annotations - participate entire group
4.

10/26/2022

Example sentences:
AllenNLP Roberta: https://demo.allennlp.org/textual-entailment/roberta-snli

- Generalized Quantifiers
    - https://docs.google.com/spreadsheets/d/1XHtCy_uLV0BnnMXIpvyTtCPdWteqJNFEqo9phlGOzEk/edit?usp=sharing
    - **Every**
    - P: Every old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: Every man is sitting in a field
    - P: Every old man is sitting in a field. H: Every man is sitting in a field
    - **Some**
    - P: Some old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: Some man is sitting in a field
    - P: Some old man is sitting in a field. H: Some man is sitting in a field
    - **No**
    - P: No old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: No man is sitting in a field
    - P: No old man is sitting in a field. H: No man is sitting in a field
    - **Not Every**
    - P: Not Every old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: Not Every man is sitting in a field
    - P: Not Every old man is sitting in a field. H: Not Every man is sitting in a field
    - **At Least two**
    - P: At Least two old men are sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: At Least two men are sitting in a field
    - P: At Least two old men are sitting in a field. H: At Least two men are sitting in a field
    - **Most**
    - P: Most old men are sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: Most men are sitting in a field
    - P: Most old men are sitting in a field. H: Most men are sitting in a field

    - **Exactly one**
    - P: Exactly one old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: Exactly one man is sitting in a field
    - P: Exactly one old man is sitting in a field. H: Exactly one man is sitting in a field
    - **All but one**
    - P: All but one old man is sitting in a field. H: A man is sitting in a field
    - P: An old man is sitting in a field. H: All but one man is sitting in a field
    - P: All but one old man is sitting in a field. H: All but one man is sitting in a field
- Negation
    - P: Two dogs are wrestling and hugging. H: There is no dog wrestling and hugging.
    - P: Nobody is riding the bicycle on one wheel. H: A person is riding a bicycle on one wheel.
    - P: A deer is jumping a fence. H: A deer isn't jumping over the fence.
    - P: A child is hitting a baseball. H: A child is missing a baseball.
    -
- Implicatives- only verbs
    - P: An old man managed to sit in a field. H: A man sat in a field
    - P: A deer is jumping over a fence. H: A deer is jumping over the fence
    - P: A wild deer is jumping a fence. H: A deer is jumping a fence
    - P: A boy is hitting a baseball. H: A child is hitting a baseball
    - P: A girl in white is dancing. H: A girl is wearing white clothes and is dancing
- Equivalence
    - P: A man is sitting in a field. H: A person is sitting in a field.
    - P: A deer is jumping over a fence. H: A deer is jumping over the fence
    - P: A wild deer is jumping a fence. H: A deer is jumping a fence
    - P: A boy likes to play badminton with his friend. H: A child played badminton with his friend.
    - P: A girl in white is dancing to her favorite song. H: A girl is wearing white clothes and dancing to her favorite song.

signature	$\beta(\text{DEL}(\cdot))$	$\beta(\text{INS}(\cdot))$	example
$+/-$	$\equiv$	$\equiv$	*he managed to escape* $\equiv$ *he escaped*
$+/\circ$	$\sqsubset$	$\sqsupset$	*he was forced to sell* $\sqsubset$ *he sold*
$\circ/-$	$\sqsupset$	$\sqsubset$	*he was permitted to live* $\sqsupset$ *he lived*
$-/+$	$\wedge$	$\wedge$	*he failed to pay* $\wedge$ *he paid*
$-/\circ$	$\mid$	$\mid$	*he refused to fight* $\mid$ *he fought*
$\circ/+$	$\smile$	$\smile$	*he hesitated to ask* $\smile$ *he asked*
$\circ/\circ$	$\#$	$\#$	*he believed he had won* $\#$ *he had won*

Table 6.4: The lexical entailment relations generated by deletions and insertions of implicatives (and nonfactives), by implication signature.

believe -> If I believe I want coffee, does not mean I want coffee - alternation
hesitated -> I hesitated to fight vs I fought - cover

			projectivity						
signature	example	monotonicity	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$+/-$	*manage to*	UP	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$+/\circ$	*force to*	UP	$\equiv$	$\sqsubset$	$\sqsupset$	$\mid$	$\mid$	$\#$	$\#$
$\circ/-$	*permit to*	UP	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile$	$\#$	$\smile$	$\#$
$-/+$	*fail to*	DOWN	$\equiv$	$\sqsupset$	$\sqsubset$	$\wedge$	$\smile$	$\mid$	$\#$
$-/\circ$	*refuse to*	DOWN	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid$	$\#$	$\mid$	$\#$
$\circ/+$	*hesitate to*	DOWN	$\equiv$	$\sqsupset$	$\sqsubset$	$\smile$	$\smile$	$\#$	$\#$
$+/+$	*admit that*	UP	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\wedge$	$\#$	$\#$
$-/-$	*pretend that*	UP	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\#$	$\wedge$	$\#$
$\circ/\circ$	*believe that*	NON	$\#$	$\#$	$\#$	$\#$	$\#$	$\#$	$\#$

Table 6.5: The monotonicity and projectivity properties of implicatives and factives, by implication signature. Some results may depend on whether one assumes a *de dicto* or *de re* reading; see text.

Discussion about:

- BLEU, METEOR and CDer
- Attachment ambiguity and parsing metrics for Natural Language to First Order Logic
-

10/25/2022

- Reading examples for Chapter 6 and Fracas - revisiting.
- With Generalized Quantifiers:

	projectivity for 1st argument							projectivity for 2nd argument						
quantifier	≡	⊏	⊐	^	|	⌣	#	≡	⊏	⊐	^	|	⌣	#
some	≡	⊏	⊐	⌣†	#	⌣†	#	≡	⊏	⊐	⌣†	#	⌣†	#
no	≡	⊐	⊏	|†	#	|†	#	≡	⊐	⊏	|†	#	|†	#
every	≡	⊐	⊏	|‡	#	|‡	#	≡	⊏	⊐	|†	|†	#	#
not every	≡	⊏	⊐	⌣‡	#	⌣‡	#	≡	⊐	⊏	⌣†	⌣†	#	#
at least two	≡	⊏	⊐	#	#	#	#	≡	⊏	⊐	#	#	#	#
most	≡	#	#	#	#	#	#	≡	⊏	⊐	|	|	#	#
exactly one	≡	#	#	#	#	#	#	≡	#	#	#	#	#	#
all but one	≡	#	#	#	#	#	#	≡	#	#	#	#	#	#

Table 6.2: Projectivity signatures for various binary generalized quantifiers for each argument position. "1st argument" refers to the restrictor NP; "2nd argument" refers to the body VP. Results marked with † or ‡ depend on the assumption of non-vacuity; see text.

10/21/2022

SICK Dataset sentence examples: - reviewed by Mihai on 10/22 and sushma to " change examples by adding all the modifiers discussed in Ch 6"

1. Forward Entailment
   P: An old man is sitting in a field. H: A man is sitting in a field
   P: A deer is jumping over a fence. H: A deer is jumping over the enclosure
   P: A wild deer is jumping a fence. H: A deer is jumping a fence
   P: A boy is hitting a baseball. H: A child is hitting a baseball
   P: A girl in white is dancing. H: A girl is wearing white clothes and is dancing
2. Reverse Entailment
   P: A man is sitting in a field. H: An old man is sitting in a field.
   P: A deer is jumping over the enclosure. H: A deer is jumping over a fence.
   P: A deer is jumping a fence. H: A wild deer is jumping a fence.
   P: A child is hitting a baseball. H: A boy is hitting a baseball.
   P: A girl is wearing white clothes and is dancing. H: A girl in white is dancing.
3. Negation
   P: Two dogs are wrestling and hugging. H: There is no dog wrestling and hugging.
   P: Nobody is riding the bicycle on one wheel. H: A person is riding a bicycle on one wheel.
   P: A deer is jumping a fence. H: A deer isn't jumping over the fence.
   P: A child is hitting a baseball. H: A child is missing a baseball.
4. Equivalence
   P: A man is sitting in a field. H: A person is sitting in a field.

P: A deer is jumping over a fence. H: A deer is jumping over the fence
P: A wild deer is jumping a fence. H: A deer is jumping a fence
P: A boy is hitting a baseball. H: A child is hitting a baseball
P: A girl in white is dancing. H: A girl is wearing white clothes and is dancing

10/20/2022

Meeting with Mihai, Eduardo, Robert, Haris and Sushma

- Concluded that C&C CCG did not really parse and failed to replace “have the right to” with “can”. Almost all of entailment, negation and equivalence remained neutral with LangPro, against the claims from the LangPro.
- We suspect the complexity of representation using Prolog could be overloaded.
- We also suspect CCG parse trees may not really be helpful in sufficient lambda derivations.
- We would like to investigate why the proof search failed.

Discussed directions:

- Alignment and compositional datasets are most reasonable directions towards Natural Logic - Mihai
- To use best of both methods: NLI with natural logic methods along with FOL with natural prover such as analytic tableau methods. - sushma. Approval from Mihai to explore and find out if such a possibility exists. To this end, come up with an example about analytic tableau method.
- Edit distance method for examples discussed by Haris.

Tasks for next week for Sushma:

- Aligner + nli_xy
- **Sick dataset - 10 examples wth modifiers with eduardo’s dataset (ch 6) allennlp**
- Work on compositional data sets.
- To read https://aclanthology.org/N03-1022.pdf
- Comments and guidance from Mihai: CCG parser is known to not work.
- When does the analytic tableau proof stop and why? From proofs search

10/16/2022 - 10/19/2022

- Testing examples on LangPro from Fracas (contd..)
- C&C CCG parser and EasyCCG parsers
- Given a CCG tree as input, worked an example for Lambda Logical Forms
- For a given CCG form, create an example with analytic tableau proof.
- Multiple Premises:
    - Downward Monotone?

        - P1: Every European can travel within Europe.P2: Every European is a person. P3: Every person who has the right to live in Europe can travel within Europe. H: Every European has the right to live in Europe.
    - Upward monotone?
        - P1: Every European has the right to live in Europe. P2: Every European is a person. P3: Every person who has the right to live in Europe can travel within Europe. H: Every European can travel within Europe.
        - P1: Each European has the right to live in Europe. P2: Every European is a person. P3: Every person who has the right to live in Europe can travel freely within Europe. H: Each European can travel freely within Europe.
- For the single premise hypothesis case, LangPro scores:
    - Forward Entailment
        - LangPro: **Neutral** P: All Europeans have the right to live in Europe as well as travel in europe. H: Some europeans can travel in Europe.
        - LangPro: **Entailment** P: An Irishman won the Nobel prize for literature. H: An Irishman won a Nobel prize.
        - LangPro: P: There are few committee members from Portugal. H: Few committee members are from southern Europe.
    - Reverse Entailment
        - LangPro: **Neutral** P: All Europeans can travel within Europe. H: Some Europeans traveled in europe.
        - LangPro: **Neutral** P: An Irishman won a Nobel prize. H: An Irishman won the Nobel prize for literature.
        - P: There are few committee members from Portugal. H: Few committee members are from southern Europe and are also from Portugal.
    - Negation
        - LangPro: **Neutral** P: All Europeans have the right to live in Europe as well as travel in europe. H: Some europeans do not travel in Europe.
        - LangPro: **Neutral** P: An Irishman won the Nobel prize for literature. H: An Italian did not win a Nobel prize.
        - P: There are few committee members from Portugal. H: Few committee members are not from Portugal.
    - Equivalence
        - LangPro: **Neutral** P: All Europeans have the right to live in Europe as well as travel in europe. H: All people in Europe have the right to live.
        - LangPro: **Neutral** P: An Irishman won the Nobel prize for literature. H: A person from Ireland did not win a Nobel prize.
        - P: There are few committee members from Portugal. H: Few committee members are Portuguese.
    -

10/15/2022

- Testing examples on LangPro from Fracas (cont..)

10/14/2022

Paper reading LangPro
To find out from Robert if the NLI_XY dataset was received.

10/12/2022

Chapter 6 (NLI)
Negation

10/11/2022

Chapter 6 (NLI)
Implications

10/8/2022

Chapter 6 (NLI)
Quantifiers

10/6/2022

Meeting with Mihai, Eduardo, Robert, Haris and Sushma

Tasks for Sushma:

1. Select 10 examples from NLI_XY dataset
2. Select examples from https://github.com/mosharafhossain/negation-and-nli and https://github.com/mosharafhossain/AFIN
3. Chapter 6 (NLI)
    a. Quantifiers
    b. Implications
    c. Negation
4. Test on the Interfaces:
    a. Langpro
    b. Nli transformer allennlp
5. Look at numbers, breaking & error patterns

6. Langpro paper discussion
7.

10/5/2022

1. The slides and literature : https://docs.google.com/presentation/d/11RnBvum2nTVz7VK_9vIPG7HgZWrzuQE0LQK9m3Tnu4M/edit?usp=sharing
2. some examples from the Colab Notebook as per some smaller examples/directions for rules/FOL
3. summarizing the "natural" LangPro paper

9/29/2022

Tasks:
Generate examples for each of the following:

- Define the task (weighted tree LSTM) - **Sushma**
- Literature for Logical Neural Networks and look for SNLI tasks - Sushma & Haris
- NLI operations on compositional dataset - NLI_XY dataset at quantifiers and formalize, plugging in the compositional operations over this dataset
- NeuroLogic Decoder - **Sushma**
- **Robert's idea:** *Maybe "verbalizing" it, something like: if someone performs in talent shows then they are engaged -* ***Robert, Sushma***
- **How many tokens that are in FOL are not in NL? - Sushma**
- **NL-Rules-FOL pipeline example - Sushma**

Challenges:

- To find convincing reason why we need formal verification just from first logic type of data
- What evidence exists that indeed first order logic or higher order logic will indeed solve formal verification of line of reasoning in a simple Natural Language Inference framework using Language Models
- The need to reduce the distance between what-is or as-is to how-to or what-next

LNNs Literature:

1. https://ibm.github.io/LNN/lnn/LNN.html?highlight=first%20order
2. https://ibm.github.io/LNN/papers.html?highlight=first%20order
3. https://github.com/IBM/AMR-CSLogic

Summary of results:
https://docs.google.com/presentation/d/1aY-idzjpGN-5NEGyCHVmAXk8UPZTpEwRQl5qsEwMJiY/edit?usp=sharing

Meeting notes (Mihai, Eduardo, Robert, Haris and Sushma) :

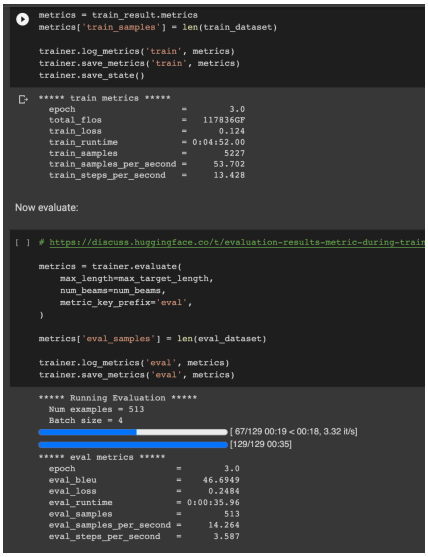
- Weighted tree LSTMs for Math equations and evaluation for formal evaluation https://openreview.net/forum?id=Hksj2WWAW&noteId=Hksj2WWAW
    - Terminal symbols use one-hot encoding
    - Each equation LHS (Left Hand Side) and RHS (Right Hand Side) is represented as an LSTM
- https://arxiv.org/pdf/2002.06544.pdf
    - Encoder decoder model by introducing a variable alignment mechanism that enables it to align variables across predicates in the predicted FOL. We further show the effectiveness of predicting the category of FOL entities - Unary, Binary, Variables and Scoped Entities, at each decoder step as an auxiliary task on improving the consistency of generated FOL. We perform rigorous evaluations and extensive ablations.
    - Lambda Dependency-based Compositional Semantics
    - They used **sequence to sequence transduction**: https://www.cs.toronto.edu/~graves/seq_trans_slides.pdf
- Siamese recurrent networks (using LSTM + GRUs) : https://arxiv.org/abs/1906.00180
- Dependency parsing for FOL and NL using Hierarchical Tree LSTMs: https://aclanthology.org/Q16-1032/
    - Start with Robert's paper: Parsing as tagging (http://clulab.cs.arizona.edu/papers/pat.pdf)
    - For parsing, predict the relative position for head
    - Input text is NL and FOL are very different (from tree structure that needs to be generated) - FOL uses predicates
    - Come up with a decoder to generate a tree
- Alignment between NL and FOL - approach to explore (similar to converting NL to SQL statements or SPARQL which have different symbols/commands)
- "X died" -> AMR is easy predicate is die
- "X kicked the bucket" -> AMR representation is complex
- Generally we do
- **Next Step:** Literature for Logical Neural Networks and look for SNLI tasks - Sushma & Haris
- Weighted tree LSTM
- Noun Phrases, verbs, adverbs, adjectives - syntax
- Build a tree between entities, verbs with FOL predicates and nouns
- Unary dependency over predicate
- Mihai's question: **"Why do we need First Order Logic for Natural Logic?"** We need first order logic for formal verification over line of reasoning, which is more intuitive to understand the line of reasoning. **- Sushma**

9/26/2022 - 9/28/2022

1. Fine Tuning the Encoder decoder model: https://github.com/sushmaakoju/folio-fol-nli/blob/main/notebooks/chap15_translation_nl_to_fol_finetune.ipynb

2. https://github.com/sushmaakoju/folio-fol-nli/blob/main/notebooks/nl_to_fol_adaptation_chap15_translation_gentle_nlp.ipynb
3. Results show that
    a. **Example Test NL statement:** People either perform in school talent shows often or are inactive and disinterested members of their community.
    Example Train Original FOL statements:
    ∀x (TalentShows(x) → Engaged(x))
    ∀x (TalentShows(x) ⋁ Inactive(x))
    Predicted FOL statement:
    ∀ x (AlentShows(x)→ ingrijireInactive(x))
    ∀ x (AlentShows(x)→ ingrijireInactive(x))

    b. **Without fine tuning vs without fine tuning:**



4. TODO for FOLIO:
    a. Weighted tree LSTMs for Math equations and evaluation for formal evaluation https://openreview.net/forum?id=Hksj2WWAW&noteId=Hksj2WWAW
        i. Terminal symbols use one-hot encoding
        ii. Each equation LHS (Left Hand Side) and RHS (Right Hand Side) is represented as an LSTM
    b. https://arxiv.org/pdf/2002.06544.pdf
        i. Encoder decoder model by introducing a variable alignment mechanism that enables it to align variables across predicates in the predicted FOL. We further show the effectiveness of predicting the category of FOL entity - Unary, Binary, Variables and Scoped Entities, at each decoder step as an auxiliary task on improving the consistency of generated FOL. We perform rigorous evaluations and extensive ablations.
        ii. Lambda Dependency-based Compositional Semantics
        iii. They used **sequence to sequence transduction**: https://www.cs.toronto.edu/~graves/seq_trans_slides.pdf

c. Siamese recurrent networks (using LSTM + GRUs) : https://arxiv.org/abs/1906.00180
d. Dependency parsing for FOL and NL using Hierarchical Tree LSTMs: https://aclanthology.org/Q16-1032/
e. Alignment between NL and FOL - approach to explore (similar to converting NL to SQL statements or SPARQL which have different symbols/commands)

9/24/2022

- **Blind Translation run over "test" NL, FOL set:**

  Implemented:
  https://github.com/clulab/gentlenlp/blob/main/notebooks/chap15_translation_en_to_ro.ipynb

  Assumption:
  1. Assume translation from EN to FOL should fail.
  2. Assume '∀', '¬', '→', '∧','∨', '(',')' to the vocabulary means something to t5 transformer, which is expected to not mean anything. (since we did not finetune to teach the transformer, so the transformer never really learned about relation between words in natural language and FOL statements).

  Approach: Add the special token in FOL (symbols)
  I.e. add all possible special tokens that are used in First order logic.
  '∀', '¬', '→', '∧','∨', '(',')' .

  Results of blind test run over T5 for NL to FOL translation:
  Source language: NL (Natural Language)
  Target language: FL (First Order Logic)
  The prediction is in german language (checked with Google Translate).

```
df = results.to_pandas()
df.loc[0,'prediction']

'<pad> Natürliche Sprache zu erster Ordnung Logik: Wenn Menschen häufig in schulischen Talentshows performieren, dann besuchen und sind
sehr engagiert bei Schulveranstaltungen.</s><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><
pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><
pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><
pad><pad><pad><pad><pad><pad>'
```

  Notebook:
  https://github.com/sushmaakoju/folio-fol-nli/blob/main/notebooks/nl_to_fol_adaptation_chap15_translation_gentle_nlp.ipynb

- **Implemented Fine Tuning tool:** - from Mihai's textbook examples which I missed checking, this was suggested by Haris:

https://github.com/clulab/gentlenlp/blob/main/notebooks/chap15_translation_en_to_ro.ipynb

- **TODO: Fine tuning using unicode strings:** '∀', '¬', '→', '∧','∨', '(',')' https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
-

9/23/2022

Error with creating converting/tokenizing using PyTorch:
https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
Revision:
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture08-nmt.pdf

9/21/2022

- .
- TODO: add analysis, metrics for this per the tutorial by wednesday
- TODO: Summarize this in pseudocode approach for the meeting:
    - "We start search for operations from source to destination and then we also simultaneously from destination backwards to source, then there exists a converging point which may be (<=n/2) (n being min depth of path from both source to dest as well as dest to source), then we can find a path such that source to some node P requires some "m" operations and from destination to P requires "k" operations, if we reverse dest to node P then we can reconstruct path from source to destination as combine(source to P , P to destination) as the shortest path. so this way we do not explore all paths. Since we converge by duality such that path is optimal and there is convergence between bidirectional due to duality in case of equivalence, forward entailment and reverse entailment. we can use bidirectional path exploration to find shortest path because duality holds true for Equivalence and Forward vs Reverse Entailment and attempt to find an approximately shortest path without seeing all possible paths but by reducing search space that holds duality".
-

9/20/2022

- Worked on encoder-decoder for NL to FOL (with symbols) using en_core_web_sm
- Wrote detailed steps and comments
- I was using PyTorch which uses SpaCY models.

- **As per suggestion from Mihai, to use Hugging Face initially before using PyTorch for some baseline.**
- Robert suggested to use T5 transformer using Hugging Face: https://huggingface.co/docs/transformers/model_doc/t5
- So implementing Hugging Face solution for NL to FOL - IN PROGRESS

9/19/2022

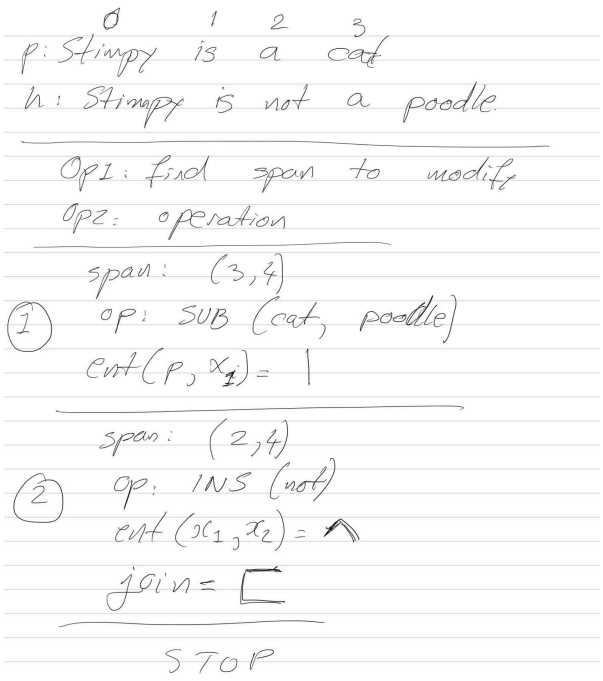
- Worked on encoder-decoder for NL to FOL (with symbols) using en_core_web_sm
- Wrote detailed steps and comments - to upload to github
- There are generally too few works towards formalizing Natural logic to First Order Logic.
- CICM conference: https://easychair.org/smart-program/CICM2022/2022-09-20.html

9/18/2022

- Wrote down a PyTorch tutorial with comments as my understanding.
- Attempting to create a dataset and modify it but it gave errors
- Fixed the errors and modified the code to use English (for just German to English)
- Sequence models from Gentle NLP textbook

9/17/2022

- Worked on Edit distance for word-wise sentences
- Worked on implementing Bidirectional Dijkstra.
- The problem : representing an edit distance matrix for sentence level in the form of a graph. What is a node and edge here?
    - Let each word in sentences be a node - if we use set, we would reduce repeated words in each sentence, but we should not. So we let the words be asis and nodes will words from each sentence. Sentence => list of words
    - Use the list but what serves as the root? # ? just that one start tag used in edit distance ?
    - Treat “tree” edit distance as “graph edit distance?
    - Parse the graph
- In the process of learning text alignment, came across: two algorithms
    - Wagner–Fischer algorithm
    - Needleman–Wunsch algorithm
    - Smith–Waterman algorithm
- But can we use NLTK with AlignSent ? https://www.nltk.org/api/nltk.align.html
- Explored this at length how to build the graph for word-wise just for



-

9/16/2022

- Z3 prover results just over 3 FOL premises with SAT/UNSAT - uploaded to github
- Explore solutions to my approach:
- “We start search for operations from source to destination and then we also simultaneously from destination backwards to source, then there exists a converging point which may be (<=n/2) (n being min depth of path from both source to dest as well as dest to source), then we can find a path such that source to some node P requires some "m" operations and from destination to P requires "k" operations, if we reverse dest to node P then we can reconstruct path from source to destination as combine(source to P , P to destination) as the shortest path. so this way we do not explore all paths. Since we converge by duality such that path is optimal and there is convergence between bidirectional due to duality in case of equivalence, forward entailment and reverse entailment. we can use bidirectional path exploration to find shortest path because duality holds true for Equivalence and Forward vs Reverse Entailment and attempt to find an approximately shortest path without seeing all possible paths but by reducing search space that holds duality”.
- ***Continued to first implement a small example for “known” set of operations at each stage:***
- ***Attempted to Implement for Edit distance algorithm with Bidirectional :*** *https://www.homepages.ucl.ac.uk/~ucahmto/math/2020/05/30/bidirectional-dijkstra.html*

9/15/2022

Meeting with Mihai, Eduardo, Robert and Haris

- Powerpoint for the presentation: https://docs.google.com/presentation/d/1pqUOw-g1nJuRcXsG2UUCQRle9IcG2HaWT6wCrec2l-A/edit?usp=sharing
- https://colab.research.google.com/drive/1xAM1BjYvsBSTxfMhMywUiUgnO07vYLbg?usp=sharing
- NLP professor working -
- **Tasks for Sushma for next week**
    - Depends on quantifiers and determiners. Study Extended WordNet - how they generated, the modifier cases - read this paper (manual rules)
    - Test Encoder-decoder model over 1) Use as-is dataset just premises and corresponding premise-FOL statements 2) Use premises and modify symbols to use “forall” “and” “or” etc instead of symbols for premise-FOL statements and show results.
    - To attempt to show Z3 prover results just over FOL premises.
- **The shortest path problem using Reinforcement Learning from word 1 to word 2:**
    - **Example : Premise:** Stimpy is a cat , **Hypothesis:** Stimpy is not a poodle
    - for the Nat Log problem you explained today - We know source = the cat and destination = not a poodle
    - We begin to find SUB/INSERT/DEL operations.
- **Exploring word-word2 edit distance paths using Duality approach - Sushma**
    - **Premise:** Stimpy is a cat , **Hypothesis:** Stimpy is not a poodle. **(Negating from independence Entailment relation. Not a dog. Not poodles.)**
    - **Premise:** Stimpy is a cat , **Hypothesis:** Stimpy is an Abyssinian.
    - **Premise:** Stimpy is a cat , **Hypothesis:** Stimpy is from a feline family. (Forward Entailment and Reverse Entailment are symmetric Duality.)
    - **Because duality holds true for Equivalence and Forward vs Reverse Entailment and attempt to find an approximately shortest path without seeing all possible paths but by reducing search space that holds duality.**
    - We start search for minimum number of operations from source to destination and then we also simultaneously start search from destination to source, then there exists a converging point which may be (<=n/2) where n is a min depth of path from both source to dest as well as dest to source, then we can find a path such that source to some node P requires some "m" operations and from destination to P requires "k" operations, if we reverse dest to node P then we can reconstruct path from source to destination as combine(source to P , P to destination) as the shortest path. so this way we do not explore all paths. Since we converge by duality such that path is optimal and there is convergence between bidirectional due to duality in case of equivalence, forward entailment and reverse entailment.
    - This is one approach: we can use bidirectional path exploration to find the shortest path. Reference:

https://efficientcodeblog.wordpress.com/2017/12/13/bidirectional-search-two-end-bfs/

-

9/14/2022

Some outstanding questions for NL to FOL conversion problem:

- What are the best representations of data for FOL statements for encoder-decoder?
- What are existing approaches to represent and/or generate data from existing FOL statements that are interpretable for encoder-decoder?
- Should the symbols remain as symbols or should they be converted to text such as ∀ be represented as ForAll, ∧⊕∨ be respectively represented as And, XOR and OR?
- Conclusion FOL was claimed to be included in the dataset, but Conclusion FOL is still missing. Should we annotate conclusion FOL manually ? would not take much time.

The question, perhaps problem to address is- what is the best way to represent a parse tree as a dataset. Some approaches I worked on deciphering are as follows:

1. We first compare the Dependency graph with that of a parser for FOL premises and generate a parse tree for FOl premises. This is what I plan to show tomorrow.
2. To attempt to show Z3 prover results just over FOL premises.
3. To discuss how to represent the Grammar and parser results in a dataset for encoding: I created a couple of FOL data representations that are better interpreted by the Encoder-decoder model vs FOL data representations in a one-hot encoding style dataset. Basically I have 3 variations for data representation for tomorrow.
4. Lastly I am working on implementing an encoder-decoder for NL to FOL and working on input/output format for the pipeline which I hope to complete in sometime.

9/13/2022

I came across a possible solution subspace within the Set vertex cover problem addressed in Combinatorics and graph theory when reading a book on Applied combinatorics.

Discussion with Robert & Haris about Entailment Relation Cover problem:

- Discussed the findings from 9/12/2022 about Set Cover problem using Combinatorics and Graph Theory but a proposed solution space to explore the Cover relation in Entailment Relations in NLI.
- **Presented a possible direction to Robert and Haris to address the problem of the Cover in NLI problem lies at the intersection of Applied Combinatorics & Graph Theory. The idea is suppose we have the "World Knowledge Graph" such as NELL: http://rtw.ml.cmu.edu/rtw/resources and parse this knowledge graph by using a solution/algorithm to Entailment Cover relation problem.**
- NLI Cover Entailment relations, animal vs non-apes - animal and non-apes have an 1) intersection of sets which is non-empty and 2) the union of the two subsets is Universal

set. The point 2 matches with the Set Cover problem i.e. to find a minimum number of subsets in the {animal, non-apes} sets.
- The Set and Graph theories do not consider “intersection” of subsets to be non-empty, as pointed out by Robert & Haris.
- However, the solution to this problem lies in a subspace of solutions or algorithms within the Set Cover problem using Graph theory solutions - either by vertex or edge cover.
- Additionally I believe that Cover in NLI needs change and/or like to find out why it was a Cover.
- The required algorithm is to address Set cover with Intersection size of at least 1.

Solutions with Set cover with Intersection size of at least 1:

- I came across this paper, could this be relevant to a solution for finding cover with a non-empty Intersection between candidate sets (example: {animal, non-apes}? https://www.sciencedirect.com/science/article/pii/0304397585902245
- **Robert’s approach (summarizing in my own words):** So for each pair of word candidates being verified for Cover Entailment relation i.e. {word1, word2} , we look for all synonyms, hyponyms and search for "Not" of word2 in word1's set of synonyms, hyponyms including word1. For example, in this case the word should be NOT (non-ape) should give ape as a result. Together this boils down to: how do we 1) get NOT(non-ape) and 2) exactly find ape in hyponyms of animal?. Apparently these two require exhaustive search since we can find antonyms of non-ape exhaustively. And somehow find ape. and we also find hyponyms of animals until we find ape. Approach to find ape : to detect not/non prefixed to the nouns. We hope to have moNLI and NLI_XY datasets. We have not yet heard from NLI_XY dataset owners yet.

9/12/2022

Findings relevant to Entailment Relations from the Book “Applied Combinatorics, by Fred S. Roberts & Barry Tessman”:

(Citation Reference book: Applied Combinatorics, by Fred S. Roberts & Barry Tessman)
Findings on Cover: Cover in Entailment Relations has an intersection in Graph Theory and Combinatorics (**Reference book:** Applied Combinatorics, by Fred S. Roberts) . There exists a Maximum Matching M such that there is a K which is a minimum cover, then |M| = |K| . If $\alpha = \{x, y\}$ and α is in M then either x or y is in K.

- More formally, Cover is a topology such that cover of a set X is a collection of sets whose union includes X as a subset.
- Set cover problem has been np-complete since 1972.
- Universe: Consider the universe U = {1, 2, 3, 4, 5} and the collection of sets S = { {1, 2, 3}, {2, 4}, {3, 4}, {4, 5} }. Clearly the union of S is U. However, we can cover all of the elements with the following, smaller number of sets: { {1, 2, 3}, {4, 5} }.
- If animal and non-ape exist such that

- This boils down to the set covering optimization problem: suppose there is a Universe U and family S of subsets of U, then the problem is there is a set of subsets k that are subsets of S that “covers” the Universe.
- Example: Bipartite graphs can be an example - A bipartite graph is a graph whose vertices can be partitioned into two disjoint sets that are independent of each other.
- This is part of a set of Problems known as Hitting Set problems.
- Hitting set problems:
    - Vertex cover
    - Edge cover
    - Geometric set cover
    - Set packing
    - Maximum coverage problem
    - Dominating set problem
    - Exact cover problem
    - Red Blue set cover
    - Set cover abduction
- The restriction of Hitting set problems, however, is constrained over “Bipartite” graphs. This necessarily means that each set (for example a pair of words that we want to verify as a cover must be disjoint and independent).
- The idea is suppose we have the “World Knowledge Graph” such as NELL: http://rtw.ml.cmu.edu/rtw/resources and parse this knowledge graph by using a solution to one of problems.

9/8/2022

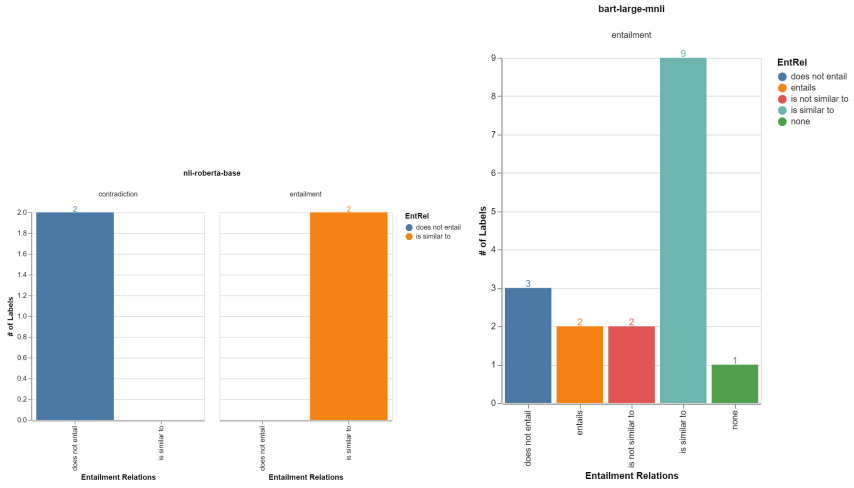
- **Meeting minutes:**
    - **FOLIO (First order logic statements) annotated dataset**
        - Directions to explore 1) NL to FOL and 2) Soft reasoner
    - NeuralLog: Natural Language Inference with Joint Neural and Logical Reasoning
        - Use of polarity indicator
        - Combining neural components with natural logic
        - Uses paraphrasing
        - Apply various transformations and select best one at each iteration
        -
    - Use MoNLI and NLI_XY (source: https://arxiv.org/pdf/2112.08289.pdf) https://raw.githubusercontent.com/atticusg/MoNLI/master/nmonli_train.jsonl
    - Under compositionality, does this still hold? Can we add concepts discussed
    - ***FOL vs Natural logic - both have tradeoffs***
    - Tasks:
        - Refer NL to FOL for dependency parsing: https://aclanthology.org/Q16-1032/
        - Can we use **NLI_XY** to add chapter 6 from Bill Mccartney? What approaches exist?
        - 30 min presentation.

- Come up with an approach to implement the encoder decoder model for NL to FOl (use Hugging face tutorial).
    - Explore the tradeoffs between the two methods.
- Contact Zheng for extracting specific tokens for first order logic - ***done***
- Adding https://arxiv.org/pdf/1412.7449.pdf .

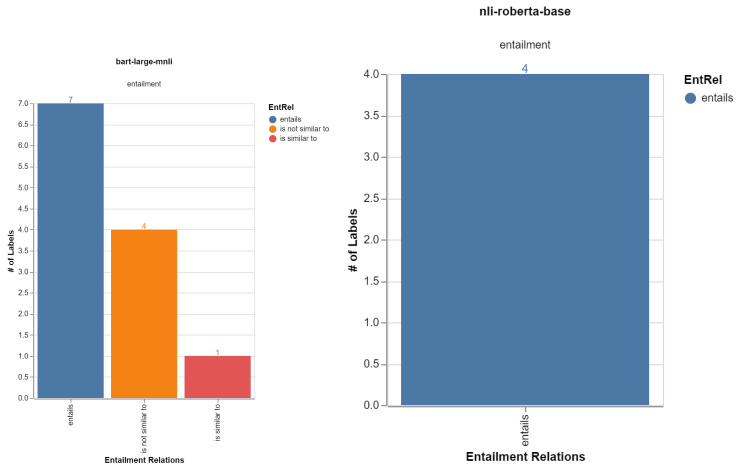
9/7/2022

- **Findings:** for few shot learning, best case would be to have following, which are missing at the moment due to limited examples from SICK and Fracas:
    - Have equal number of examples for each Entailment Relations
    - For each entailment relation, to have positive and negative examples
    - To use smaller batch size shttps://arxiv.org/pdf/1412.7449.pdfuch as 1 instead of 32 since examples are too few
    - To work/modify with variations in hyper parameter settings (learning rate: 2e-05)
- **Results:**
    - We consider only Equivalence, Negation, Forward Entailment, Reverse Entailment.
    - Fracas & SICK dataset counts per Entailment Relation
        - Note: None category was an attempt to get negative examples as opposed to positive ones for each of the Entailment Relations.

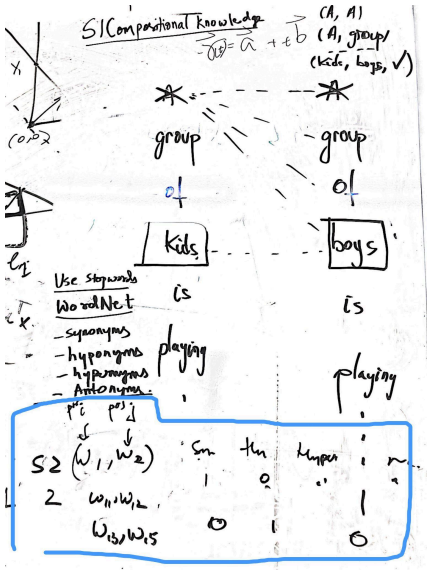
Dataset	# FE	# RE	# Equv	# Neg	# None	# Total
Fracas	4	0	1	4	0	9
SICK	2	2	8	2	1	15

    - Fracas sentences extracted and annotated over Roberta Base NLI & BART Large NLI models
    - 

    - SICK sentences extracted using WordNet and manually annotated over Roberta Base NLI & BART Large NLI models



-
    - Ran Few shot learning over Roberta Base NLI and too few examples for both SICK and Fracas
    -
- Challenges :
    - Too few examples to test over Few shot
    - Fracas dataset most sentences were outside of Entailment relations (Projection & Join operations over Entailment Relations.)
    - To select sentences from a sick dataset was a bit challenging.
    - Spent more time trying to extract data from SICK using WordNet
    - *Manually labeled SICK dataset examples.*
    - Accuracy is not included
- Implemented for zero-shot learning for both Fracas and SICK datasets.
- Modified to only compare words that have the same Parts of speech to avoid infinite recursion or longer, lengthier search results to halt the search over limited.
- To use one hot encoding & to use stopwords, indices of word 1 & 2 from sentence -> suggested by Robert



- I formally discussed with Robert about my implementation and the above suggestion to use one-hot encoding was his suggestion.

9/6/2022

Meeting with Mihai:

- Use adjectives, adverbs, nouns and verbs using WordNet for 4 Entailment relations: Eqv, Neg, FE, RE
- Use Fracas and SICK datasets (10 examples each for Ent Rels for SICK)
- SICK - Sentences Involving Compositional Knowledge: https://huggingface.co/datasets/sick

9/5/2022

Fracas Dataset Comprehension

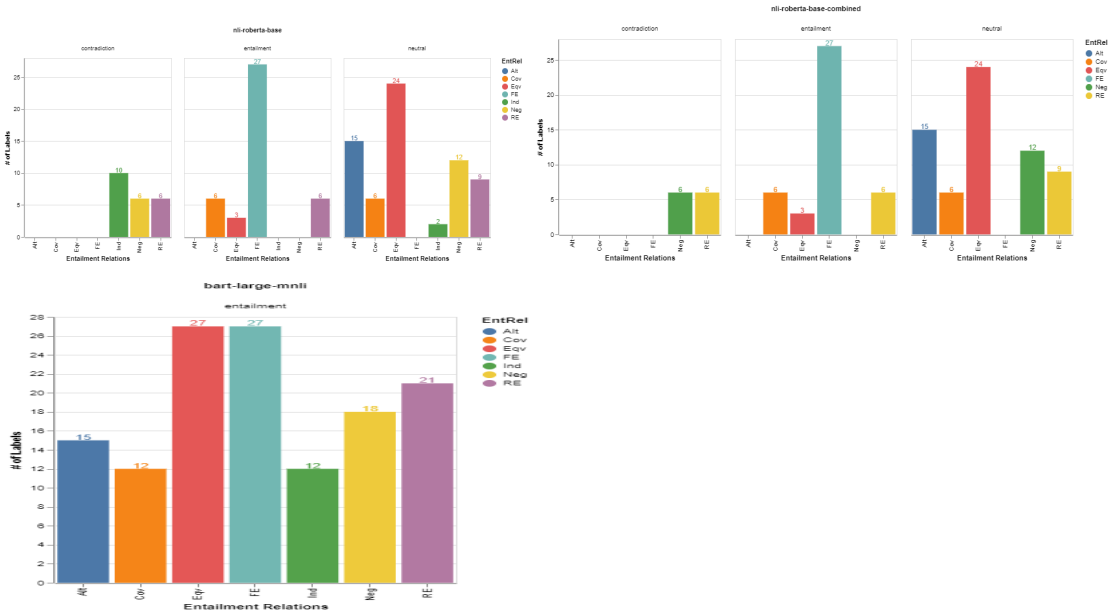
- Defining Entailment relations with that sentences from Fracas Dataset
    - The dataset is classified into following categories:
        - 1 GENERALIZED QUANTIFIERS
        - 2 PLURALS
        - 3 (NOMINAL) ANAPHORA
        - 4 ELLIPSIS
        - 5 ADJECTIVES
        - 6 COMPARATIVES
        - 7 TEMPORAL REFERENCE
        - 8 VERBS
        - 9 ATTITUDES

- The full categories and subcategories from Fracas dataset is as follows:
    - 1 GENERALIZED QUANTIFIERS
      Q is a set of Quantifiers.
        - 1.1 Conservativity
            - For a given Determiner D, D(A,B) <-> D(A, A ∩ B).
        - 1.2 Monotonicity (upwards on second argument)
            - Right upward monotone: For a g and g' such that g ⊆ g', Q ∈ D, if X in Q and X ⊆ Y, then Y in Q
            - if P≤Q then Det N P≤Det N Q
        - 1.3 Monotonicity (downwards on second argument)
        - 1.4 Monotonicity (upwards on first argument)
        - 1.5 Monotonicity (downwards on first argument)

9/1/2022

Mihai, Eduardo, Robert, Sushma, Haris

Analysis of Zero shot learning:



Model	Expected Input	Modified Input	Expected output	Actual Output
nli-roberta-base	Premise & Hypothesis sentence pairs	S1 + S2, W1 ent_rel W2	Entailment, Contradiction and neutral	Except for the Forward Entailment examples which are correctly labeled as Entailment, all others are scattered as per above plot, as expected with base model.
nli-roberta-base	Same as above	S1 + S2, W1 ent_rel W2	Entailment, Contradiction and neutral	Same as above, no improvement.
bart-large-nli	Same as above	S1 + S2, W1 ent_rel W2, W1 not_ent_rel	Is entailment_rel, not	all sentences were labeled as Entailments correctly.

		W2	entailment_relation	**Pitfalls:** BART Large NLI model seems to generalize well, which also seems too good to be true. Missing Counterfactuals i.e. we need negative examples for each of the Entailment relations.

Feedback:

- restrict the scope of the Sentence examples to an existing dataset on NLI or Entailments.
- As a next natural step, we use Fracas dataset to select sentence pairs that match the Entailment Relations: https://nlp.stanford.edu/~wcmac/downloads/fracas.xml :
    1. **Equivalence:** Two synonyms for Equivalence
    2. **Forward Entailment:** Hyponym for FE (Forward Entailment)
    3. **Reverse Entailment:** Hypernym for RE (Reverse Entailment)
    4. **Negation:** For negation, we need to come up with examples that match no intersection + Universal set. Possibly antonyms that match the Set theoretic representation.
    5. **Alternation:** Use different synsets with Hyponyms as root from Wordnet.
    6. **Cover:** First find the antonym of the word from Sentence 1 for example. Then select an antonym, find the hyponym of the antonym - parsing the WordNet for the word selected from Sentence 1, as an example.
    7. **Independence:** select any two sentences that do not fit into any of above cases - requires more evaluation.
    8. Sentences from Fracas dataset

    a. The output would be yes/no.

    b. We conduct manual evaluation as the evaluation approach

    c. Use relation text such as “entails” instead of “is entailment of”

    d. Additionally, report accuracy for the next meeting.

    e. About the examples:
        - We consider generating negative examples for each of the Entailment relation sentence pairs from Fracas dataset task listed above.
        - For each of the counter cases, we generate and consider such a sentence that suggests the corresponding Entailment Relation is not true.

    f. Other ideas
        1. Idea suggestion by Robert: since there will be labels for the sentence, maybe something from Multi-Instance Learning can be helpful? For the future

Todos:

1. **Sushma** to select sentence pairs for each of the Entailment relations from the Fracas Test set and evaluate for entailment relations using the approach provided in above sections.
2. **Robert** to review and provide feedback on the sentence examples shared by Sushma and we share this with Mihai
3. **Sushma** to test the selected sentences over zero shot learning and include accuracy metrics and more analysis.
4. **Robert** to evaluate and provide feedback before the next Thursday meeting for zero shot analysis.
5. **Sushma, Robert, Haris, Mihai and Eduardo: To read following papers for next week:**
    1. https://yanaiela.github.io/papers/winograd-square-one.pdf
    2. Faithful Reasoning Using Large Language Models

8/31/2022

Preparatory notes for the meeting, follow up tasks (Sushma):

- **Questions:**

Why are x and y denoted as elements (small case) as opposed to sets in NLI dissertation?

- **Challenges:**
    1. There seem to be no compositionally annotated datasets over Entailment Relations yet.
    2. We define sentence examples over Entailment Relations using word level entailments for simpler “Atomic” cases.
    3. For word level approaches, explored a couple of directions and Robert suggested using WordNet with a specific graph traversal for auto-generating sentence-level examples for Entailment Relations.
    4. However the sentences were constructed using Wiktionary and WordNet was not used. Mihai guided and reviewed.
    5. The Fracas Test set from https://nlp.stanford.edu/~wcmac/downloads/fracas.xml has specific cases generalized to and .
        a. For example, the problem set has categories broadly addressing Generalized Quantifiers, Plurals, Anaphora, Ellipsis etc.
        b. This set does not contain or suggest Entailment Relations at compositional level.
        c. Besides, the problem set consists of Premises (p, q including question), hypothesis (h, declarative hypothesis), Answer(a, yes/no/don't know),

why (justification and/or explanation of the answer), fracas_answer('yes', 'no', 'unknown', 'undef').
    d. Thus this task requires compositionally annotated dataset to further extend or test the Entailment Relations teach, train and test Deep Learning models.
6. Came up with some wrong examples, corrected as per initial review from Mihai.

- **Defining the 7 Entailment Relations with examples using Sets for simpler word-level entailments: Reference: NLI Dissertation by [Bill McCartney]**
    1. **Equivalence ≡:** $x \equiv y$ iff $x = y$. Simpler case over which Equivalence is holds: Let A and B are two sets such that
        a. $\exists x,\ x \in A\ or\ x \subset A$ and
        b. $\exists y,\ y \in B\ or\ y \subset B$
        c. $\forall x,\ \forall y,\ x \rightarrow y$
        d. Example: couch ≡ sofa
    2. **Forward Entailment ⊏:** $x \sqsubset y$ iff $x \subset y$. Simpler case over which Forward Entailment holds: Let A and B are two sets such that $\exists X,\ X \subseteq A$ and $\exists Y,\ Y \subseteq B$.
        a. $\forall X,\ \forall Y,\ X \subset Y$
        b. then $X \sqsubset Y$
        c. Example: crow ⊏ bird
    3. **Reverse Entailment ⊐:** $x \sqsupset y$ iff $x \subset y$. Simpler case over which Reverse Entailment holds: Let A and B are two sets such that $\exists X,\ X \subseteq A$ and $\exists Y,\ Y \subseteq B$.
        a. $\forall X,\ \forall Y,\ X \supset Y$
        b. then $X \sqsupset Y$
        c. Example: asian ⊐ thai
    4. **Negation ^:** $x \wedge y$ iff $x \cap y = \phi$ and $x \cup y = U$. Simpler case over which Negation is holds: Let X and Y be two sets such that
        a. $X \cap Y = \phi$ **and**
        b. $X \cup Y = U$
        c. $\forall X,\ \forall Y,\ X \wedge Y$
        d. Example: able ^ unable
    5. **Alternation |:** $x \mid y$ iff $x \cap y = \phi$ and $x \cup y \neq U$. Simpler case over which Negation holds: Let X and Y be two sets such that
        a. $X \cap Y = \phi$ **and**
        b. $X \cup Y \neq U$
        c. $\forall X,\ \forall Y,\ X \mid Y$
        d. Example: cat | dog
    6. **Cover ⌣ :** $x \frown y$ iff $x \cap y \neq \phi$ and $x \cup y = U$. Simpler case over which Negation holds: Let X and Y be two sets such that
        a. $X \cap Y = \phi$ **and**
        b. $X \cup Y \neq U$
        c. $\forall X,\ \forall Y,\ X \frown Y$

    d. Example: Animal $\smile$ None-ape
7. **Independence # :** $x \# y$ iff $x \cap y \neq \phi$ and $x \cup y = U$. Simpler case over which Negation holds: Let X and Y be two sets such that
    a. All other cases
    b. $\forall X, \forall Y, X \# Y$
    c. Example: Hungry # Hippo

- **Approach followed for example sentences for each of the 7 Entailment Relations:**
    1. Selected words from Wiktionary to select words that are hyponyms, hypernyms, synonyms, antonyms
    2. Words selected were from
        a. Synonyms for *Equivalence* for preserving Symmetry
        b. Hyponyms for *Forward Entailments*
        c. Hypernyms for *Reverse Entailments*
        d. Antonyms for *Negation*
        e. No x y is a Negation of The x y
        f. General Real World entities for Alternation, Cover and Independence
    3. Selected few other synonyms from Thesaurus.com
    4. Verified most of the selected words in WordNet (as per suggestion from Robert)
    5. Depending on appropriate matches found for each Entailment relation, I selected one method over the other and listed that as the source for each of the examples.
    6. For simplicity, we just show word to word entailment relations for now.
    7. We construct simple sentences which do not require JOINs or projection operations and/or do not prefer phrases or sentences that require JOIN & Projections over Entailment Relations.

Sentence Construction

1. First we select a pair of words that hold the representing Entailment Relation from the selected words, for creating a pair of sentence examples.
2. We then construct the sentence that possibly preserves the representing Entailment Relation between the two words that were selected.
3. We evaluated (manually) each of the pairs of words towards Set Theoretic representation for each of the Entailment Relations the word pairs represent.
4. Pending review from Mihai and NatLog - 9/1/2022

The Example sentences created are as follows:

1. The example sentences written have 9 sheets.
2. First sheet formal definitions of each of Entailment relations
3. The 7 sheets represent examples for 7 Entailment Relations.
4. For each pair of sentence examples, we select 2 words that hold the representing Entailment Relation.
5. The last sheet attempts to show types of “Entailments” (Not Entailment relations).

6. For each of the Entailment Relations, there are two sentences and two words, such that from sentence S1, there exists one of the words selected towards Entailment Relations the example represents.
7. Similarly for Sentence S2, there is word W2 that represents the word-level Entailment relations with word W1.
8. Lastly there are two additional columns for the 7 groups of examples, Source column is the actual source that contributed towards selecting the words for constructing sentences.

Example Dataset:

https://docs.google.com/spreadsheets/d/1pSmDl6WVP5RXApW3AB6dLz6NrCk5Y8NjX9fRRwwsLUQ/edit?usp=sharing

Zero Shot Learning with Open Prompt Configuration over Natural Language Inference Models:

The models used for evaluating zero-shot learning over Example sentence pairs towards Entailment Relations are as per Seniors from the lab (Robert, Zheng):

1. NLI-roberta-base
2. Zero-shot-classification using the pipeline

https://colab.research.google.com/drive/134GynijAbPGzx1aArQgz6tHKJw6-OCMB?usp=sharing

The following example is done in addition to the aforementioned two models:

1. Bart NLI

8/25/2022

- Mihai, Eduardo, Haris & Sushma: Entailment relations from NLI Dissertation.
- Followup Questions & tasks for next meeting:
    1. Is there a dataset annotated compositionally for entailment? SNLI dataset. -*Sushma*
    2. If not, can you set up a zero-shot evaluation using texts from Fracas (or something similar) as inputs? - *Sushma*
    3. Show an example of a prompt for each of the 7 basic relations. - *Sushma*
    4. Evaluation on 3 dimensions:
        a. basic vs. compositional relations
        b. base or large models
        c. zero shot vs. few shot (start with zero)