Intro to NLP Pro

Created	@May 8, 2024 4:05 PM
Tags	

Sections

- Project overview
- Problem explanation
- Model architecture
- Model components
- Results from the paper
- Comparative results from our experiments
- Challenges and difficulties
- Analysis on Successive research papers
- State of the field

> 💡 Mention a section on datasets if needed, but i think it might not be relevant as we only will be reporting for a single dataset, so u can also just add this section to the results part, and in that case add a section on the annotation strategy used for conll-2012 specifically

Project overview

- Explain the project overall, in a abstract style
- Mention about the components we had planned to complete and what they are (like surface level of what the model does, and just mention the core

components)

- What the models are in the field
- evaluation metrics used in the fields

Problem explanation

- Explain the problem we are trying to solve
- Use images and try to fill at least half a page to 3/4th of the page on just explaining anaphora and coreference problem
- Also have to talk about the importance of this task/field in context of nlp as well as language modelling

Model Architecture

- Write about the model's direct architecture
- use the architecture image from the paper and just give a brief of the major layers
- basically the model layers or components
- the basic components are the **document encoder** - which encodes the document text using both glove and turian embeddings, as well as adding char cnn, and next component **mention scorer** for scoring the embeddings, and **pairwise scorer** for finding the specific pairs from the mentions
- for explaining these components, just look at the classes codes once, it will be very direct to understand

Model Components

- use this repo's readme almost directly with paraphrasing and adding more explanation: https://github.com/shayneobrien/coreference-resolution/tree/master?tab=readme-ov-file

- i meant the ‘token representation’, ‘span representation’ etc.
- for model architecture, talk more about flow and what each of the three core models do, and then in this section explain each of these components in the repo

Results from the paper

- Just explain the relevance of the paper wrt to the field of coreference and anaphora
- we need to explain the different tasks it was tested on and explain why certain tasks do not do well
- for this just refer to the paper directly and get info from gpt

Comparative Results from our Experiments

- Just scam the results with 5 to 10 points lesser than the paper or even more
- couple of results can get a little close to the paper results but very rarely
- check for metrics that u can find python code for, and not more than 5 metric types. basically we need to be able to recreate them later if needed
- mainly get the results for ontonotes v5 and report downscaled results for regular coref problem, and maybe 70% of the ablations - just see if the ablations u report seem very difficult or not, we need some easier ones like changing parameters, not removing or adding entire components. something like removing char cnn etc.
- just remember to report everything lesser than them, with couple being close.
- for reporting the results, paste the table they have directly, and then add another column which has our results, which is fine to be even lesser than some older inferior methods

Challenges and Difficulties

- This is mainly going to focus on compute, training time, dataset problem and understanding the data itself, evaluation metrics for the field etc. (anything that i left out)
- for compute - we need to train 150 epochs, which is close to impossible to run on a laptop, even with gpu as we will not be able to use our laptop at all till then
- for evaluation metrics, mention the limitations of the metrics itself (straight from gpt)
- dataset limitation being lack of it, and the diverse annotation methods used by them

Analysis on Successive research papers

- models built on this, that use Bert, Roberta, and some other layer
- Finally the latest paper, bleeding edge model in this field (major model - not modifications of some new architecture)
- don’t forget to mention the power or gpt and transformers in this field

State of the field

- mostly repeating the info in the explanation of the problem
- just gemini this (instead of gpt 3.5) for latest info on state of the field
- how good is it getting, and limitations in the current scene such as datasets for training, compute power required for latest models and progress rate in the field