M20Temp17: Advanced NLP **Monsoon 2020**

Course Project and Term Paper Logistics

Instructor: Dr. Manish Shrivastava *TA: Mounika Marreddy, Prashant Kodali*

1 Course Project Logistics

- It will be a group project, with every team consisting of two members.
- A group can choose a project from the list below, or propose a project of their own, while meeting the criteria elaborated in the class. If you propose your own project, it is subject to review and approval of faculty. Even if you are proposing your own project idea please fill in project preferences from the list too.
- Project will be allocated on first come first serve basis.
- Each project will have a mentor. It is the project group's responsibility to reach out the mentor and making continuous progress over the course of semester.
- With project descriptions, baseline and baseline + are indicated. Baseline is the bare minimum that is expected out of the project. Baseline + lists the improvements that can be carried out by you. A team is in no ways limited by the literature listed in project description. A good literature survey and its implementation, with due consultation and review of faculty / mentor will go a long way in project evaluation.
- There will be three submissions for this project.
  1. Project Outline Submission: A week after project allocations. You will have to give a gist of the project (description, datasets), literature survey, plan for project execution.
  2. Interim Submission: Report on the project progress. This will be in the middle of the semester. Date will be communicated along with project allocations.
  3. Final Submission : Project Report, Presentation and Code submissions.

2 Term Paper Logistics

Term paper is directly linked with your Course Project. A literature survey and thorough critical analysis of various techniques used for that problem will have to be carried out. This effort will also help in your course project.
Evaluation: Presentation at the end of semester, along with Project Presentation.

3 Project Domains

- Machine Translation
- Question Answering
- Summarization
- Conversational Systems

4 Projects List

1. **Domain:** Machine Translation

   Description: ACL 2019 shared task - Machine Translation of News Articles

   http://www.statmt.org/wmt19/translation-task.html

   Dataset: Pick one of the language pairs of your choice from the list in the link

   Baseline: Explore and Implement project

   Basline+: Over and over the base implementation explore and implement the improvements.

2. **Domain:** Question and Answering

   Description: Implement Rank QA : Neural Question Answering with Answer Re-Ranking paper. Andcompare with other state-of-art papers of your choice. Also compare your results on different datasets like: SQUAD , WIKI

   https://www.aclweb.org/anthology/P19-1611.pdf

   Dataset: SQUAD , WIKI

   Baseline: Implement the paper on SQUAD and WIKI data sets project

   Basline+: Compare with other state-of-art papers Question and Answeringpapers of your choice(minimum 2)

3. **Domain:** Summarization

   Description: Scientific Document Summarization shared task

   https://github.com/WING-NUS/scisumm-corpus

   Dataset: Link above

   Baseline: Explore and Implement project

   Basline+: Over and over the base implementation explore and implement the improvements.

4. **Domain:** Machine Translation

Description: Given a paragraph/document in english wiki, 'translate' it into simple wiki - Should work even for docs out of wiki.

http://www.cs.pomona.edu/~dkauchak/simplification/(Aligned Data)

Dataset: Link above

Baseline: a) Statistical MT Baseline; b) Encoder-Decoder/Word Embeddings Approach

Basline+: Over and over the base implementation explore and implement the improvements.

5. **Domain:** Question and Answering

   Description: Implement a state-of-art paper on Community Question and Answering and Propose your improvements over the baseline model

   Dataset: Yahoo Answers

   Baseline: Implement a state-of-art paper of your choice for baseline

   Basline+: Suggest improvements over the baseline and implement them.

6. **Domain:** Machine Translation

   Description: Domain Term Extraction

   Paper: Term extraction using non-technical corpora as a point of leverage

   Dataset: Prepare Dataset by web scraping Wikipedia

   Baseline: Collect Data from Wikipedia and implement a model (either from one of the papers or a hybrid model) . The goal is , on a new document , we should be able to identify the Domain Terms

   Basline+: Improve the Data , Come up with suggestions on your Baseline to improve the results.

7. **Domain:** Question and Answering

   Description: Implement State-of-art papers on Open Domain Question and Answering

   Dataset: WikiQA dataset

   Baseline: Implement a paper of your choice as the Baseline model.

   Basline+: Explore more papers in the same area and come up with an improved model of your baseline model

8. **Domain:** Machine Translation

   Description: To incorporate the benefits of multiple MT systems into one, so as to improve upon the performances of the individual baseline systems.

Variants: 1. HMM; 2. Attention based models

Approach 1 (HMM):

(a) MVP

- For a given bilingual corpus, train a translation system using Moses (or any online PBSMT API will do too, in case Moses is hard to set up).
- Translate the test corpus, and run BLEU to score the translation.
- Similarly, using OpenNMT or Tensor2Tensor (any NMT framework), train an attention model, and run it on the test corpus. Again, run BLEU to score the translation
- Compare both the scores obtained.
- Obtain a monolingual corpora (maybe even the target side of the bilingual corpora) of the target language, and train a language model (preferably with KenLM)
- For each translated sentence of the test corpus (using both PBSMT and NMT), evaluate the perplexities of both using the LM, and pick the better translated sentence. (Also find percentage of choosing translations of both systems.)
- Again, run BLEU to score this newly created mixed set of translations.
- Observe improvements in BLEU scores, if any.

(b) Steps Forward

- For each translated sentence of the test corpus (using both PBSMT and NMT), use both as observation sequences, and train a HMM model that predicts the hidden sequence. (Here, the hidden sequence implies a final sentence, which captures the best of both the translations.)
- Observe improvements in BLEU scores on the set of final sentences, if any.

Approach 2 (Attention-based models):

(a) MVP

- For a given bilingual corpus, train a translation system using Moses (or any online PBSMT API will do too, in case Moses is hard to set up).
- Translate the test corpus, and run BLEU to score the translation.
- Use a Deep Learning Framework (pytorch or keras probably) to code up an attention model. Train this model on the bilingual corpus, and run it on the test corpus. Again, run BLEU to score the translation.
- Compare both the scores obtained.

(b) Steps Forward

- Implement the Neural System Combination for Machine Translation paper (https://arxiv.org/pdf/1704.06393.pdf)

- Essentially, feed the pre-translations (already obtained in the above steps) of both systems as inputs, and run an attention on each to obtain 2 context vectors. Now, run another attention on these 2 context vectors, to obtain a final context vector, which then goes through the conventional decoding step. This model outputs a final sentence, which captures the best of both the translations.
- Train this model, and run it on the outputs obtained previously. Again, run BLEU to score the new translations.
- Observe improvements in BLEU scores, if any.

Data Resources

(a) SMT Readings

- PBSMT slides: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/SNLP/SMT

(b) NMT Readings

- Seq2Seq: https://arxiv.org/pdf/1409.3215.pdf
- Bahdanau's Attention: https://arxiv.org/pdf/1409.0473.pdf

(c) English-Hindi Corpora:

- IIT-B Parallel Corpora:`http://www.cfilt.iitb.ac.in/iitb_parallel/`
- ILCI Parallel Corpora: Will be given later.

(d) Pre-Processing Tools:

- English: Git clone the Moses decoder repository, and you will find everything needed under the scripts/tokenizer section. Refer here for instructions on how to run the scripts. (Check out the Corpus Preparation section.)
- Hindi: Git clone the IndicNLP repository, and you will find everything needed under the src/indicnlp section. Make sure to first export the appropriate path variables. Refer here for instructions on how to run the scripts.

9. **Domain:** Argument Mining

   Description: Argument Mining: Detect Arguments and claims in unstructured data

   Dataset: Link

   Baseline: Sequence Labelling on Essays Dataset

   Basline+: Relation prediction (for/against) between premises and claims.

10. **Domain:** Bias Detection in News

    Description: Bias Detection in news articles. Detect sentence level and article level bias in news domain.

    Dataset: SemEval 2019 dataset on news bias

    Baseline: Hierarchical Attention Nets for Document Classification

Basline+: Pretraining Evaluations with Sentence Subjectivity classification. Language Models, remove publisher annotation bias from training

11. **Domain:** Semantic Textual Similarity

    Description: Semantic Textual Similarity: To address the problem of semantic coincidence between sentence pairs. Commonly knownas paraphrase identification.

    Dataset: Semeval STS, SICK, MSRPC

    Track1

    Baseline: Siamese Recurrent Architectures for Learning Sentence Similarity + Augment word representations with character level features

    Basline+:
    Further modifications of your own, or
    Some suggestions
    - Trying out different similarity functions Adding a cross attention layer
    - Check out DCN

    Track2

    Baseline: : Bilateral Multi-Perspective Matching for Natural Language Sentences [with any one of the 4 matching layers under section 3.2]]

    Basline+: Further modifications of your own or Add Attentive Matching layer

12. **Domain:**

    Description: Natural Language Inference : Tounderstand semantic concepts like textual entailment and contradiction. The task isthat of comparing two sentences and identifying the relationshipbetween them.

    Dataset: SNLI, SICK Datasets

    Track1

    Baseline: Structrued Self-attentive Sentence Embedding

    Basline+:
    Further modifications of your own, or
    Some suggestions
    - Augment word representations with character level features
    - Check out DCN

    Track2

    Baseline: : A Decomposable Attention Model for Natural Language Inference

    Basline+: Add intra-sentence attention - referenced in the paper

13. **Domain:** NLP for Social Media
    List of Projects:
    (a) Sentiment analysis of code mixed data
    (b) Hate speech detection on (Gab.ai,Voat.co,Reddit)
    (c) NER detection in code-mixed data