Learning Sentence Representations from Explicit Discourse Relations

Ujwal Narayan
20171170

Saujas Vaduguru
20171098

October 9, 2019

1 Overview

Developing an accurate representations of sentences efficiently is one of a core problem in NLP. In this project, we propose to exploit discourse relations in order to better represent sentences. The intuition behind this is the fact that discourse relations annotate deep conceptual relations between sentences and these can be leveraged in order to make sure the model understands the sentence more clearly. For this we use the discourse prediction task. Consider the following sentence: *"She's late to class _____ she missed the bus"*. A human would fill in the word *'because'*. This prediction of the discourse marker is the discourse prediction task that we propose to use to train a model to represent sentences by harnessing the information in these discourse markers.

2 Goals

We aim to complete the following tasks as part of this project:

1. Automatically extract a corpus for the discourse prediction task from text data, using dependency parsing, using the method proposed in [Nie et al., 2017]
2. Create sentence embeddings for Hindi by training on the discourse prediction task
3. Evaluate the sentence embeddings so obtained on a down-stream task, namely sentiment analysis
4. *Extra Credit:* Implement a sentence encoder model and train it to represent sentences using the discourse prediction task

3 Data and Models

We propose to use the following datasets for the project:

1. **Hindi Corpus**: We will use the Parallel Hindi English Dataset curated by IIT Bombay [Kunchukuttan et al., 2018] to train the sentence embedding. The dataset is available at http://www.cfilt.iitb.ac.in/iitb_parallel.
2. **Hindi Discourse Treebank**: We will use the Hindi Discourse Treebank available at https://github.com/aarjay00/discourse_parsing to evaluate the method for automatic extraction of the discourse prediction task dataset.
3. **Sentiment Analyis**: We will use the Aspect based Sentiment Analysis corpus [Akhtar et al., 2016] curated by IIT Patna for the task of sentiment analysis. The dataset is available at http://www.iitp.ac.in/~ai-nlp-ml/resources/shad/ReviewSentimentDataset-AspectTermExtraction.zip.
4. **Sentence Encoder Model**: We use the implementation of the *InferSent* sentence encoder model [Conneau et al., 2017] available at https://github.com/facebookresearch/InferSent.

4 Milestones

- **First evaluation**: Implement automatic extraction of corpus, and validate the extraction using the Hindi Discourse Treebank.
- **Second evaluation**: Train the sentence encoder model on the discourse prediction task.
- **Final evaluation**: Evaluate the the sentence encoder model on the chosen downstream task. If time permits, implement a sentence encoder model and report results by training with that model.

References

[Akhtar et al., 2016] Akhtar, M. S., Ekbal, A., and Bhattacharyya, P. (2016). Aspect based sentiment analysis in hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

[Kunchukuttan et al., 2018] Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi Parallel Corpus. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Nie et al., 2017] Nie, A., Bennett, E. D., and Goodman, N. D. (2017). Dissent: Sentence representation learning from explicit discourse relations.