INLP Project Outline

**Team Number:** 58
Members:
Abhijeeth Singam – 2019101065
Rishabh Khanna – 2019113025
Tushar Jain – 2019101091

Coreference Resolution

In this project, we focus on a specific form of coreference resolution which is character identification in multi-party dialogues. This project is based off of the SemEval 2018 Task 4 [1]. Our aim is to compare the performance of the winning model of the Shared Task (2018) with the current SOTA architecture.

The task is to determine which character from the show 'Friends' is referenced by the different referents in the dialogues of the show
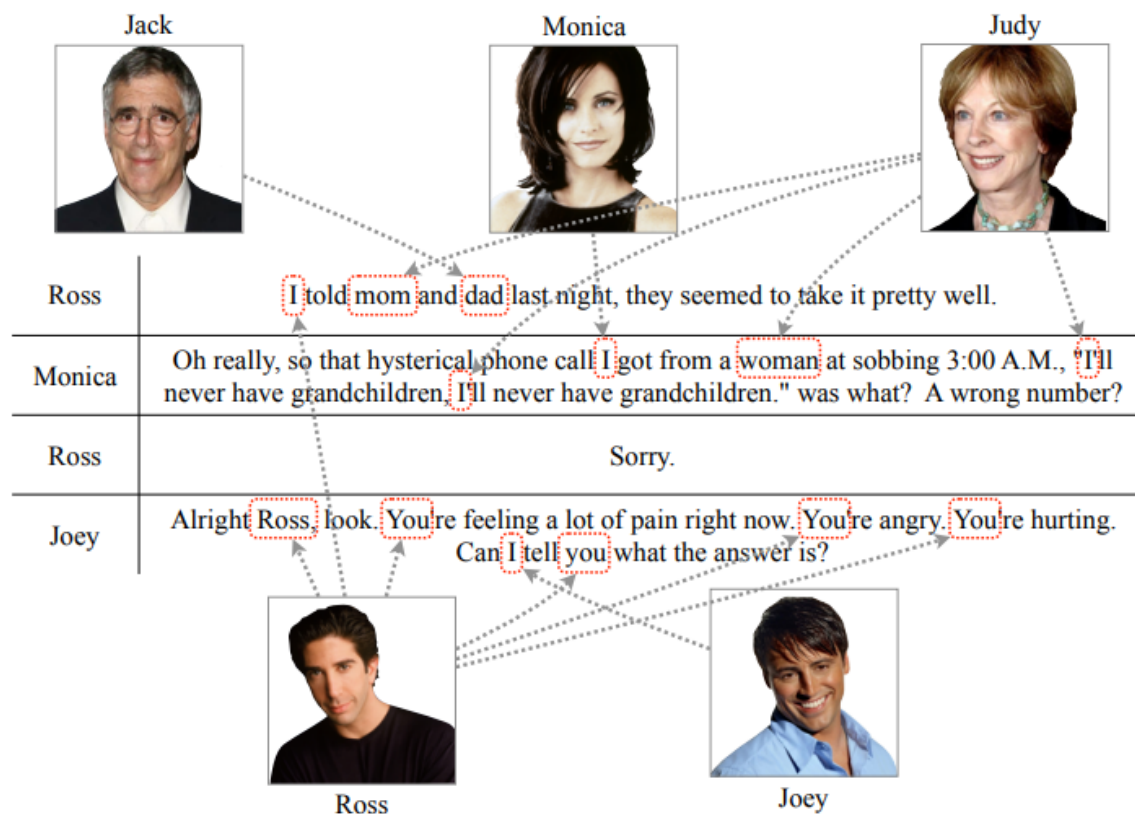
Datasets

We use the dataset used in the shared task, which is provided by the organizers on their github page [1].

The first two seasons of the TV show Friends are annotated for this task. Each season consists of episodes, each episode comprises scenes, and each scene is segmented into sentences. Each sentence is annotated with which referent refers to which character of the show.

The dataset consists of 16K sentences and 132K tokens.

Sample Data:

```
/friends-s01e01  0  0    All     DT      (TOP(S(NP*     all    -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  1    right   NN      *)             right  -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  2    ,       ,       *              ,      -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  3    c'mere  RB      (ADVP*)        c'mere -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  4    ,       ,       *              ,      -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  5    gim     VBP     (VP*           gim    -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  6    me      PRP     (S(NP*)        I      -   -    Phoebe_Buffay    *    (292)
/friends-s01e01  0  7    your    PRP$    (NP*           you    -   -    Phoebe_Buffay    *    (248)
/friends-s01e01  0  8    feet    NNS     *)))           foot   -   -    Phoebe_Buffay    *    -
/friends-s01e01  0  9    .       .       *))            .      -   -    Phoebe_Buffay    *    -
```

The last column indicates which character was referred through a numeric token.

Literature Review

SemEval 2018 Task 4: Character Identification on Multiparty Dialogues [3]
This paper describes the task, which is a shared task to perform character identification on the dataset. It describes the top two scoring systems of this shared task. The two systems described are AMORE-UPF System and the KNU-CI system.

The AMORE-UPF system approaches this task as a multi-class classification. It uses a bidirectional LSTM. This LSTM takes the input and compares them to vectors in the LSTMs hidden state. This comparison then resolves mentions.

The KNU-CI system tackles this task as a sequence-labeling problem. It uses an RNN encoder-decoder model. An RNN has poor performance for long sequences, and they solve that by applying an attention, position encoding, and the self-matching network to the original RNN.

Anaphora and coreference resolution: A review [2]
This paper is a survey paper of the current methods of anaphora and coreference resolution as of 2020 and would serve as our main source of guidance and exploration when trying to understand the task and other forms of the task at hand. We would also use this paper to find suitable baselines and alternate approaches for our task.

Project Timeline

Problem Domain Exploration + Literature Review (25th Feb - 5th March)
Shared Task winning model implementation (6th March - 20th March)
SOTA model implementation (21st March - 10th April)
Model Comparison and Evaluation (11th April - 20th April)

References

1. Shared Task GitHub Link: https://github.com/emorynlp/semeval-2018-task4
2. Anaphora Resolution Survey Paper:
   https://www.sciencedirect.com/science/article/pii/S1566253519303677
3. https://aclanthology.org/S18-1007.pdf