Eugene

Week 9 - Coreference Resolution

Introduction

- Coreference Resolution is the process of identifying all mentions of the same entity.

- We can use it for full text understanding, machine translation, and dialogue systems.

- It is a difficult problem. See Winograd Schema for example (https://pdfs.semanticscholar.org/94c0/630d50e351dc7f0bfd4ebd323053bae673f9.pdf).

- We can break the problem down into two steps: detect the mentions and cluster the mentions. The former is easy, but the latter is hard.

- For mention detection, we can use POS tagger, NER system, and constituency parser.

Linguistics Terminologies

- Coreference is when two mentions refer to the same entity in the world.

- Anaphora is when it refers to another term and the interpretation of it is determined by the context.

- Not all anaphora is coreference.

- Cataphora is referred to something only mentioned in the future.

Coreference Model

MENTION PAIR AND RANKING MODEL

- Coreferent in three simple steps:

```
1. Mention Detection
2. Coreferent calculation for every pair
3. Add coreferent link if score is high
```

- Coreferent score can be calculated with cross entropy.

- It is unable to process long sentences.

- To calculate the coreferent score:

```
1. We can use features such as semantic compatibility, gender
agreement, syntactic constraints, etc...
2. We can use a standard feed-forward neural network with word
embedding and features as input and score as output.
```

END-TO-END COREFERENCE RESOLUTION

- Based on state-of-the-art model by Lee et. al. (github.com/kentonl/e2e-coref).

- Joint mention detection and clustering with no need for pre-processing.

- Need to pick a subsequent to prune the number of pairs.

- Reason over all possible spans and assign an antecedent to every span.

- Model can assign credit to the mentions or antecedent factors.

- Start with word embedding, feed it to a bidirectional LSTM that finds the span head from a span representation, then return the mention score.

- Choice of embedding might cause issues.

CLUSTERING MODEL

- Perform agglomerative clustering (or hierarchical clustering).

- Agglomerative clustering is a hard problem because it is an EXPSPACE problem.

- Train a Mention-Pair Encoder to get a Mention-Ranking Model, and train a Cluster-Pair Encoder to get a Cluster-Ranking Model.

- Mention-Ranking Model builds the bottom of the tree.

- Cluster-Ranking Model decides how to build the tree (typically using reinforcement learning).

- We can use distance from antecedent to mention, proximal related words, part of speech, and other useful stuff for Mention-Pair Encoder features.

- Mention-Ranking Model loss function is defined in this paper Kevin et. al. (https://arxiv.org/pdf/1606.01323.pdf).

- Cluster-Ranking Model would decide if we should merge a pair of clusters, which can be done using reinforcement learning (https://cs.stanford.edu/~kevclark/resources/clark-manning-emnlp2016-deep.pdf).