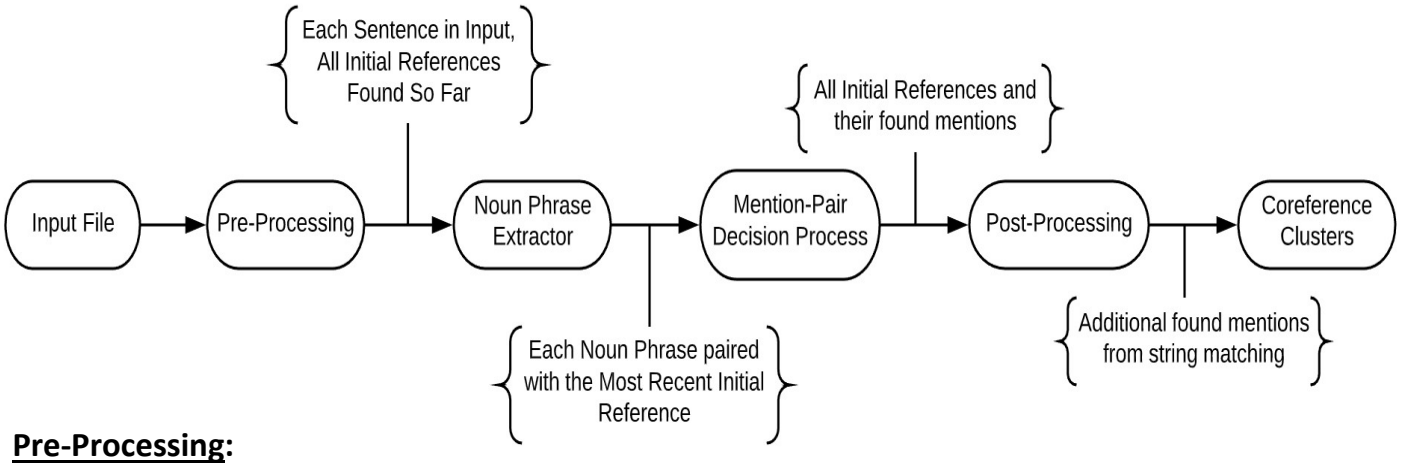
# LINGUISTICS NUT

THE UNIVERSITY OF UTAH

**Jiahui Chen**

**F1:** 0.478 **Recall:** 0.409 **Precision:** 0.574

## Overview:




### Pre-Processing:

All sentences are extracted and indexed.
Initial references from each sentence are extracted and added to a stack of initial references.
Noun phrase extraction and mention-pair classification are done on one sentence at a time.

### Noun Phrase Extractor:

Noun phrases found through spaCy's syntactic parsing.

### Mention Pair Decision Process:

Each NP is paired with the nearest initial reference preceding it [1]. The below checks occur to for mention pair classification. If mention pair is found, then NP is added to cluster of initial ref.

#### Synonym Check:

Both head nouns of the phrases in the pair get their noun synonym sets from WordNet.
If any synonym from either set match, then the pair is classified as a mention pair.

#### String Matching:

If any word in one phrase of the pair is a substring of any word in the other phrase, then the pair is classified as a mention pair.

### Tools:

`spaCy` – python package with many NLP functions and pre-trained deep learning models

- `en_core_web_lg`: largest available English language model (CNN)
- Word vectors for similarity calculation
- POS tagging for noun phrase extraction
- NER
- Lemmatization, to check for matching plurality

`NLTK` – another NLP python package, only used for access to WordNet

- WordNet used to get sets of synonyms

#### Word Vector Similarity:

The spaCy package (see Tools section for more details) assigns word vectors to most common English words. The cosine similarity between the averaged word vectors of both phrases in the pair is taken.
If the similarity score is above *0.80* (this similarity threshold was found via cross validation), then the pair is classified as a mention pair.

#### Features:

When no synonym or string match is found AND spaCy has no word vector for any word in each phrase, my own features [1, 2] are extracted for the pair and their weighted combination is used as a similarity score.

**Plurality Match:** count of words that have matching plurality (matching lemmas) in pair
**NER Match:** count of matching NERs in pair
**Capitalization Difference:** count of word capitalization differences in pair

```
manual_sim = 0.65 + (plurality(p1, p2) * 0.7) + (0.7 * ner(p1, p2)) - (0.1 * cap_diffs(p1, p2))
```

### Post-Processing:

All noun phrases are removed from the sentence, and any leftover words that string match initial references are added to coreference clusters.

**Other (Less Successful Attempts):** Mention pair machine Learning models where the Mention-Pair Decision Process is currently used: Decision Tree and SVM [1].

### Sources:

[1] "Improving Machine Learning Approaches to Coreference Resolution" Vincent Ng and Claire Cardie
[2] "Coreference Resolution" Slides from Chris Manning, Roger Levy, Altaf Rahman, Vincent Ng, Heeyoung Lee