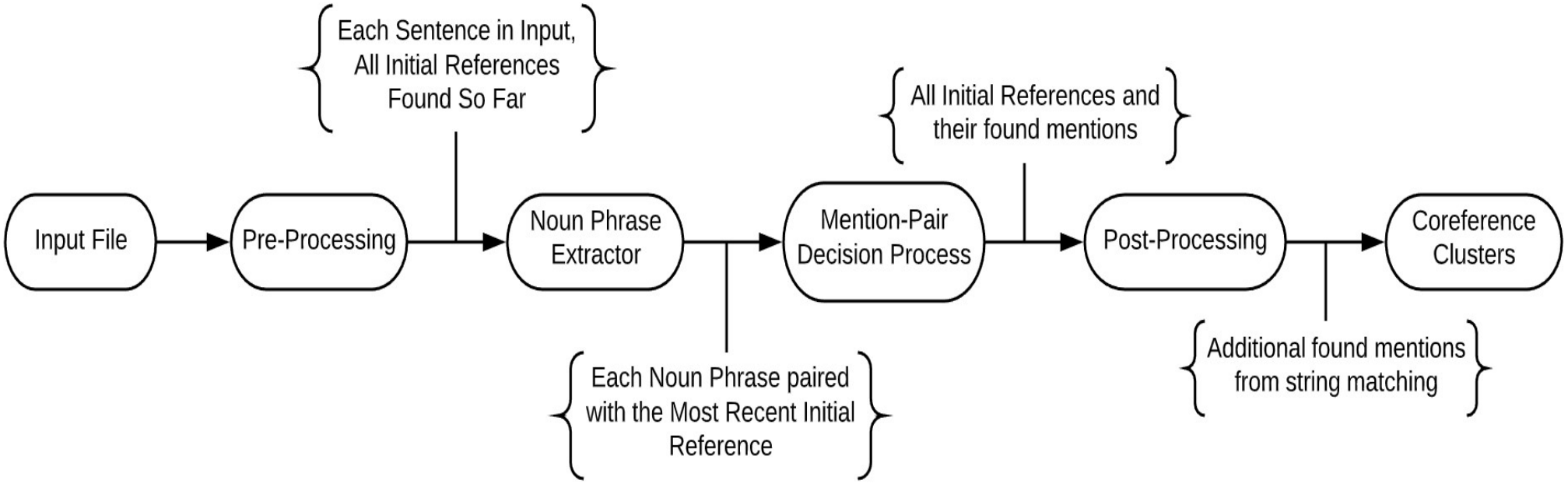
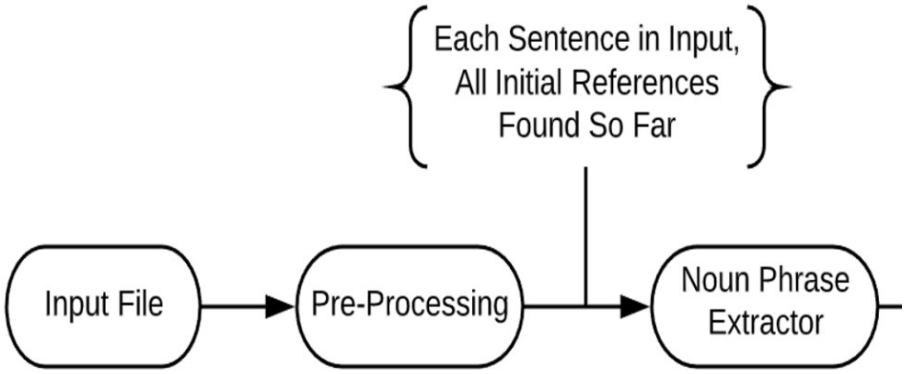
Linguistics Nut

Jiahui Chen

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





Tools Used:

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

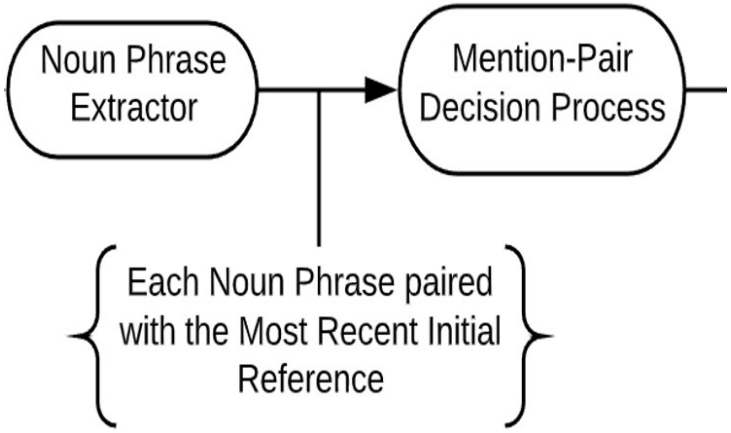
- POS tagging for noun phrase extraction

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.



Tools:

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

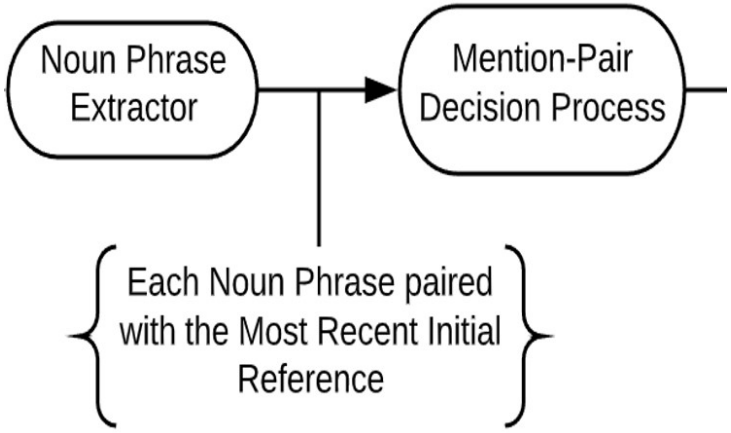
- WordNet used to get sets of synonyms

Mention Pair Decision Process Overview:

Each NP is paired with the nearest initial reference preceding it ("best match" pairing proposed in [1]).
The following checks occur on the pair 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.



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 assigned by `en_core_web_lg`

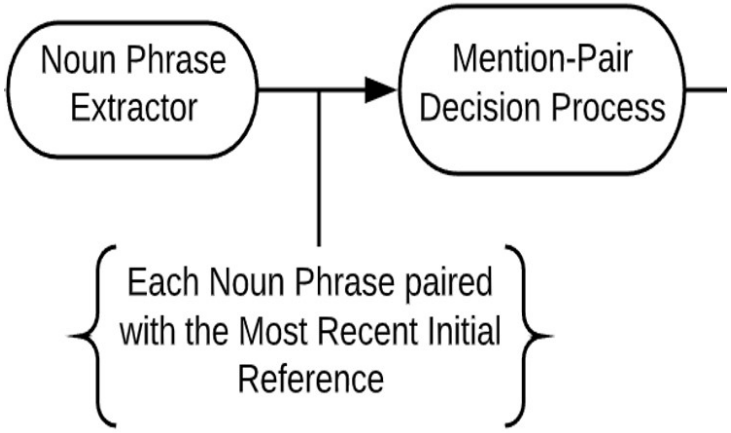
String Matching:

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.

Word Vector Similarity:

The spaCy package's model assigns word vectors to most most English words. The cosine similarity between the averaged word vectors of both phrases in the pair is taken.

If this cosign similarity is above *0.80* (this similarity threshold was found via cross validation), then the pair is classified as a mention pair.



Tools:

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

- NER
- Lemmatization, to check for matching plurality

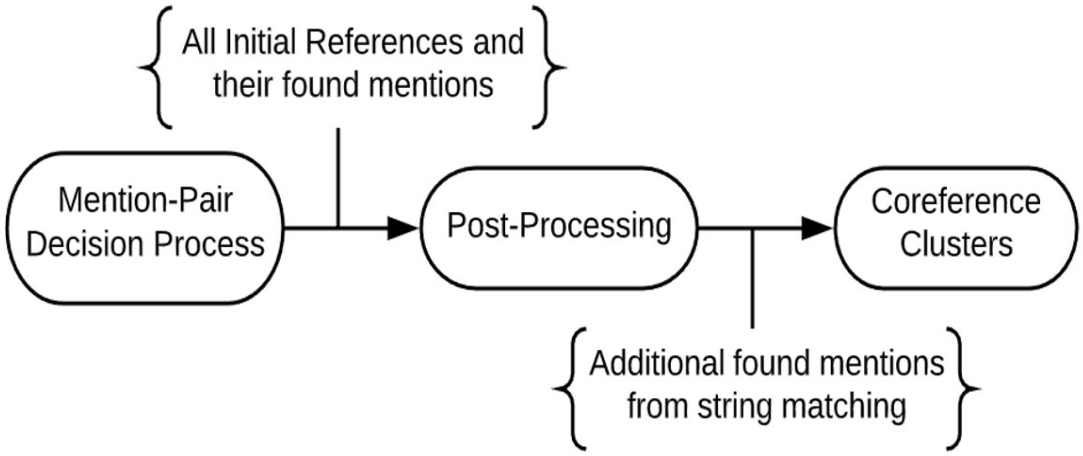
Features:

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 (considered mention pair if above 0.80).

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

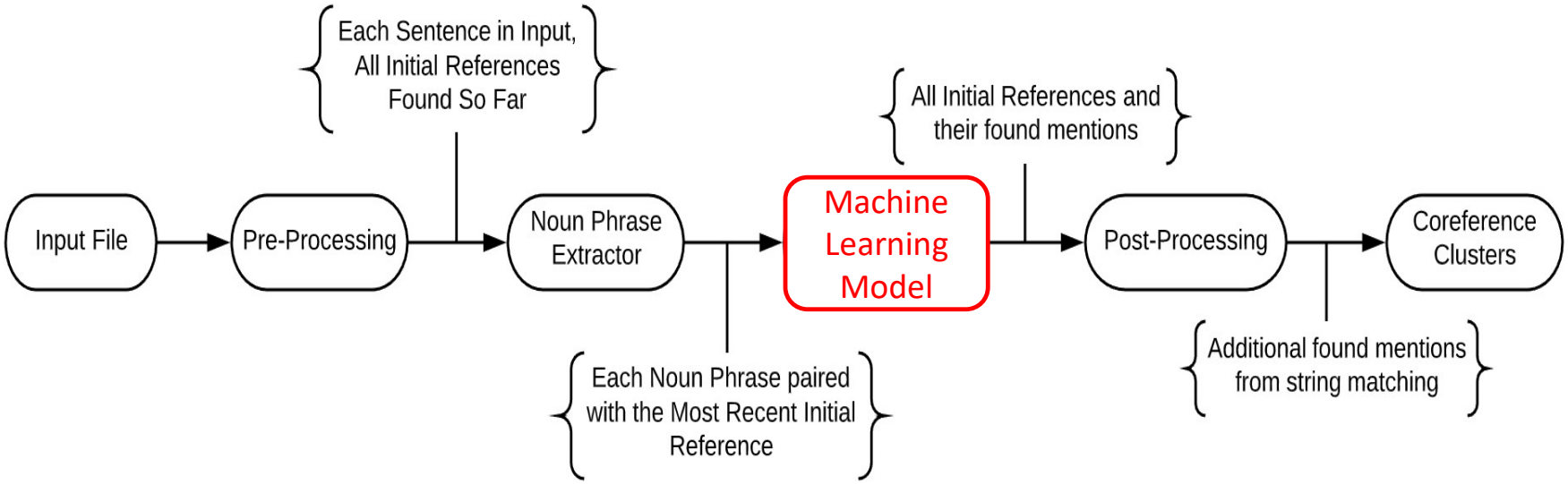


Post-Processing:

All noun phrases are removed from the original sentence, any remaining 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)



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