VT27

SRIVATHSAN G R ; NIKHIL RAMESH

THE UNIVERSITY OF UTAH

COREFERENCE RESOLUTION SYSTEM

DATA PRE-PROCESSING

- Dictionary is Created for coreferences within the document with <CoRef,Sid> as key.
- Regular expressions employed to clean the sentences and extract the corefs , designed specifically for the given dataset.

COREF DICTIONARY CREATION

- An additional Dictionary for corefs is created using wordNet to hold synonyms, it's lemmatized forms and hyponyms.
- For ex., wages and salary belong to the same Coref cluster. This dictionary helps to deal with these cases.

SENTENCE PARSING

- Designed a customized regex based parser to return the Noun phrases in a sentence.
- 8 types of Regex was applied on the tree returned by the NLTK parser to identify different types of the maximal noun phrases in the document.
- Example: {<DT>?<JJ>*<IN>*<NN>*<IN>*<NN>+}

COREFERENCE MATCHING

- A smart comparison function designed to remove stop words, comparing only the head nouns based on the POS tags.
- Each Coref word is compared with the NP sentence chuck generated by parser.
- Pronouns are handled separately using NER.
- Coreference pertaining to date are resolved by converting it to common format before comparison.

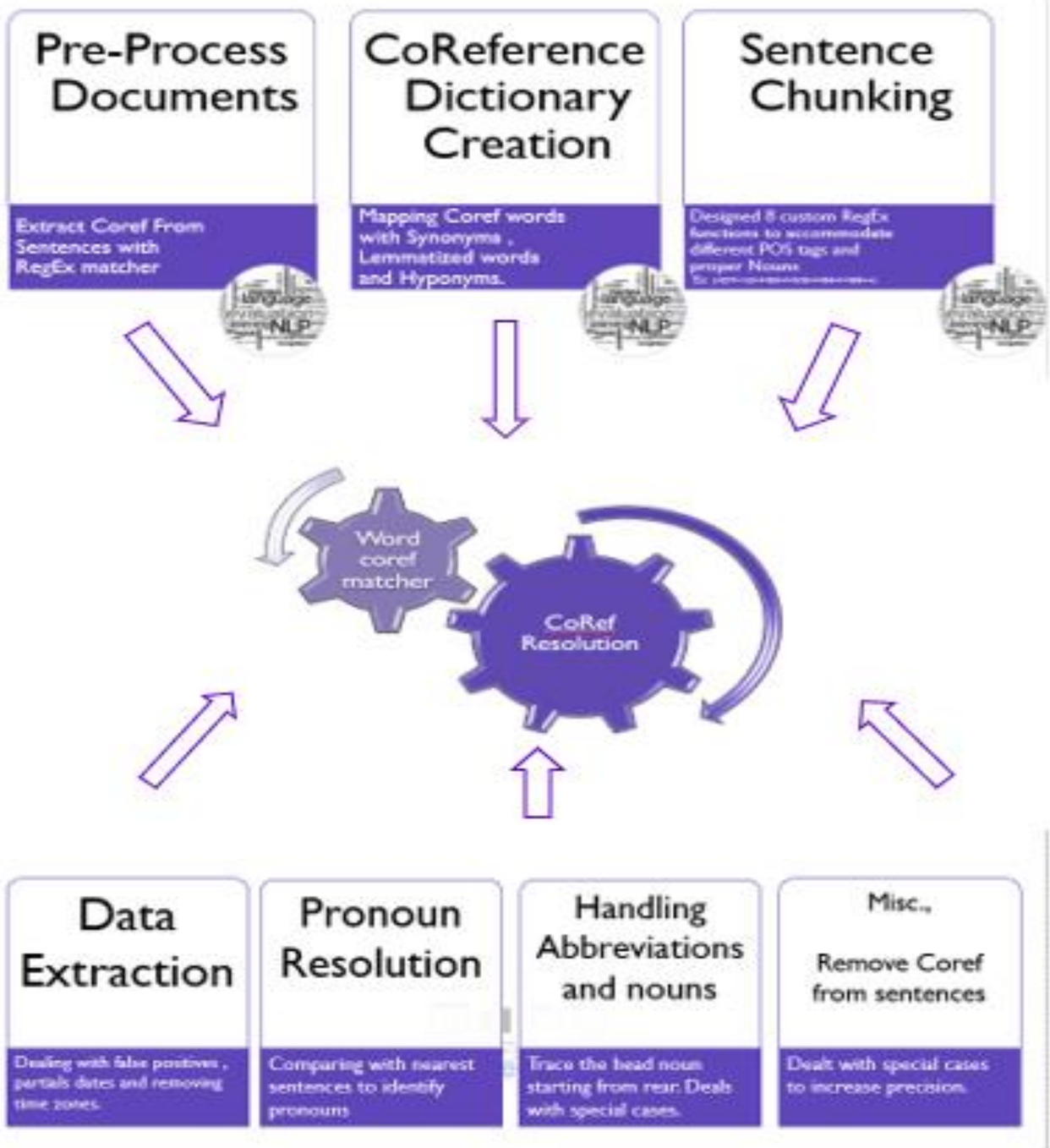
WHAT WORKED WELL

- Chunking of different types of NP in the noisy dataset, extracting the maximal phrases.
- Abbreviations and Synonyms matching.
- Eliminate timezone from dates for higher precision.
- Remove Coref from sentences.

CHALLENGES

- Dealing with CAPS sentences
- Wrong POS tagging by NLTK parser.
- Identifying head noun in Noun Phrase.

SYSTEM ARCHITECTURE



SYSTEM ANALYSIS

Pros	Cons
• Smart Sentence chunking; Handles complex noun phrases perfectly	• Large number of False positives due to reliance on WordNet synonyms.
• Accurate date resolution	• Handling noise and sentences with all caps

RESULTS

F-SCORE	**0.44548**
PRECISION	**0.48191 (333/691)**
RECALL	**0.41418 (333/804)**

CONCLUSION

- The Coreference resolution project showed us the challenges faced when NLP tasks are deployed in real world. Challenges such as dates, ambiguous sentence formations, Semantic errors were tricky to handle and are specific to the noise in the dataset.
- The system developed was able to robustly handle a large number of the aforementioned challenges and was able to balance precision and recall to improve the F-score

EXTERNAL RESOURCES

- NLTK (www.nltk.org)
- Spacy
- DateConvertor