A², Coreference Resolution System, CS6340, Fall 2016

Avani Sharma & Aishwarya Asesh

School of Computing, University of Utah

SYSTEM DESIGN AND COMPONENTS

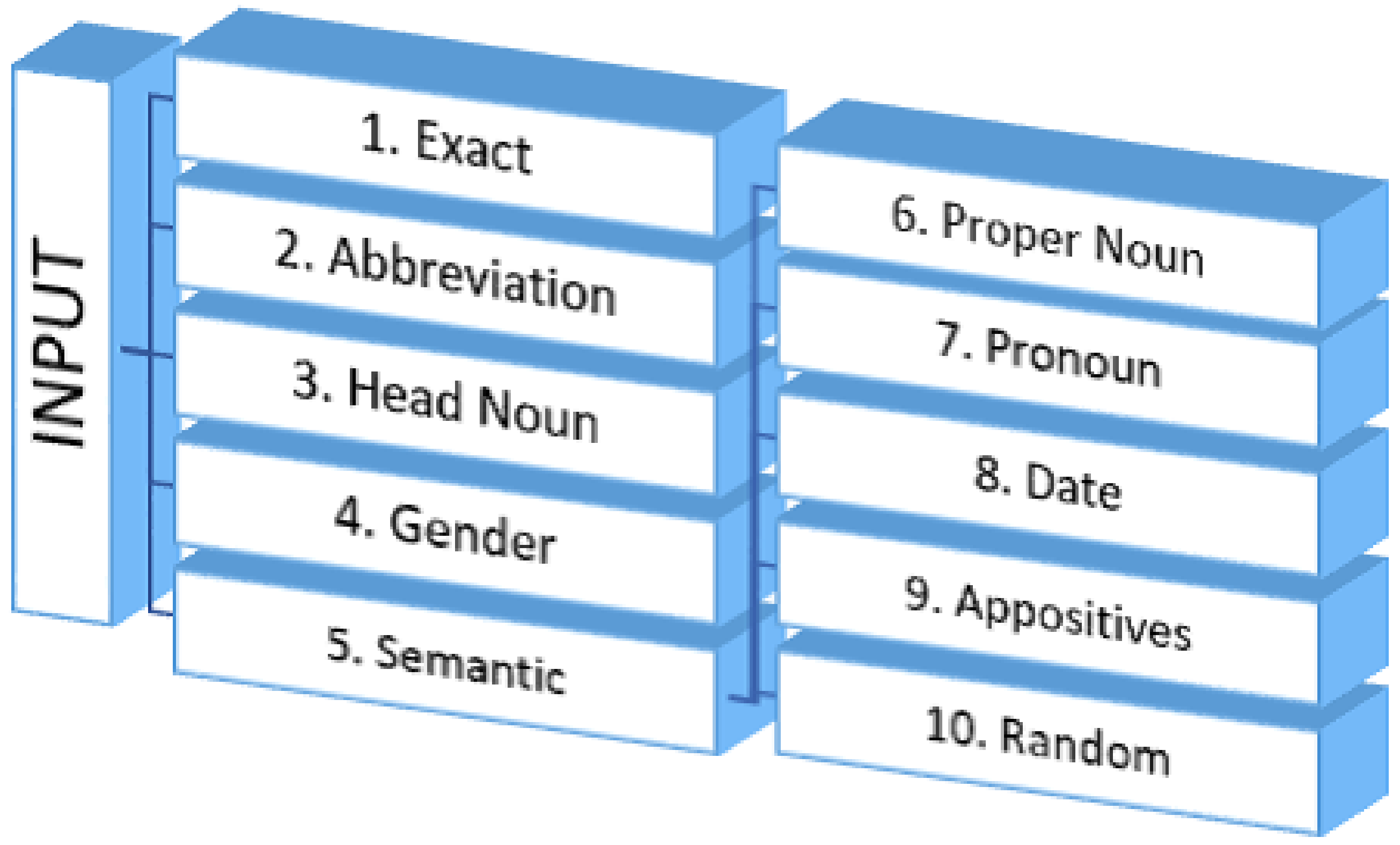

Fig. 1: Overall Basic System Structure

COMPONENTS EXPLAINED

1. Word referred with its **exact** occurrence in the document.
2. Shortened word/phrase referred with same word or order.
3. Reference given if **partial** match of a string exists.
4. Reference exist if male/female pronoun found in the list.
5. Reference made if word found in same **semantic class**.
6. If Capitalized word matches animate/inanimate pronouns.
7. If pronoun type is same, then a reference is recorded.
8. Match if pattern follows **(RE)** regular expression format.
9. If two noun phrases are **adjacent**, then Appositives.
10. Remaining values referenced to nearest preceding value.

Did you know?
The lack of a proper coreference resolution system is a hinderance in development of **Super Intelligent Robots**. Super AI consistently fails while making coreference resolutions such as understanding what **- it** refers in sentence - He saw the doughnut on the table and ate **it**. This is the reason AI specialists are serious about **Elon Musk's** prediction of potentially dangerous encounter with super intelligent silicon.

INTRODUCTION

Coreference Resolution is the task of finding, whether two or more expressions in a text base, refer to the same object or living entity.

Coreference Resolution System is an essential step for many high order Natural Language processing implementations such as summarization from a paragraph, building a question answering system, and cases were we need to do information extraction based on available database. Such systems are a mix of concepts including **NLP, Info Retrieval, Info Extraction, ML, Knowledge Representation, Logic and Inference, Sematic Search**.

For this particular project, we are more concerned in coreference resolution in cases where domain of search space is limited to a particular document.

Emphasis/Originality
Key Observation: Variation in accuracy rates occurs due to change of order for exact, substring, abbreviations, gender and other matching methods. The order for matching was finalized after analysing experimental results.

Team Member Contribution:
Aishwarya Asesh - Abbreviations, Pronoun, Date, Appositives, Random Matching.
Avani Sharma - Exact, Head Noun, Gender, Semantic, Proper Name Matching.

Guess the ***It***
Case I: This Sharp TV's picture quality is so bad, our old Sony TV is much better.**It** is also so expensive.
Case II: This Sharp TV's picture quality is so bad, our old Sony TV is much better.**It** is also more reliable.
The Case I (It) refers to Sharp, Case II (It) refers to Sony.

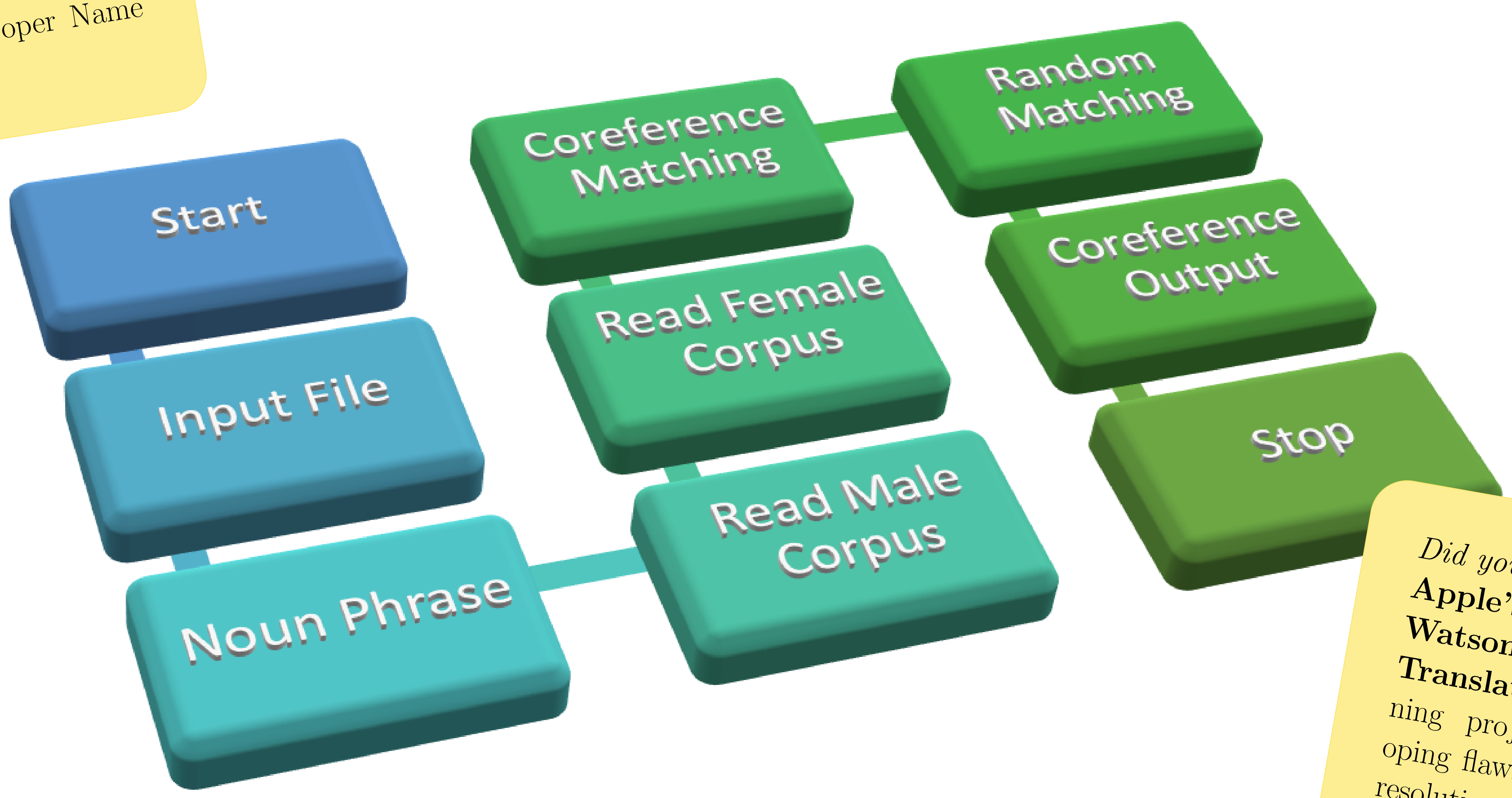

Fig. 2: Process for Coreference Resolution System

Did you know?
Apple's Siri, IBM's Watson and Google Translate all have running projects for developing flawless Coreference resolution system. People are curious about the data driven or hybrid approach used in IBM Watson, namely the most advanced Coreference Resolver. The development will be made public in the coming years.

SECOND THOUGHTS

Successes - Things that went well

1. The Exact, Substring & Abbreviation Match are the key strengths of our algorithm.
2. We are most proud of the **generalized implementation**, which makes our system balanced irrespective of the scope of the given input file.

Regrets - Scope for improvement

1. Due to **restrictions of external libraries** available for python, we were not able to get proper results during implementation process.
2. Usage of Wordnet for semantic category decisions did not yield perfect results.
3. **Stanford Core NLP** Toolkit appeared to give better results than NLTK toolkit.
4. We restricted ourselves to python toolkits only.

Did you know?
Paper titled **"Corpus-Based Identification of Non-Anaphoric Noun Phrases", by Bean, D. and our instructor Riloff, E.** is one of the earliest papers that describes a modern day approach to the problem of Coreference Resolution.

EXTERNAL RESOURCES

1. Thanks to our friend NLTK, Wordnet and lxml.
2. Stopwords for Head Noun/Substring match taken from www.nltk.org/book/ch02.html
3. Male/Female corpus used from www.cs.cmu.edu/Groups/AI/util/areas/nlp/corpora/names/0.html
4. Noun Phrase Coreference as Clustering , Cardie and Wagstaff, EMNLP 2000
5. A Multi-Pass Sieve for Coreference Resolution , Raghunathan et al., EMNLP 2010
6. Deterministic coreference resolution based on entity-centric, precision-ranked rules , Lee et al., Computational Linguistics 39(4), 2013

PERFORMANCE

Due to the generalized Implementation strategy used, our system performed in a very balanced manner
Evaluation Results: TEST SET 2 = 60.46%, TEST SET 3 = 61.11%

LaTeX TikZposter