INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

Coreference Resolution for Hindi

Objectives

Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

- It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.

Example

Who can forget the legendary opening batting pair of Sachin Tendulkar and Sourav Ganguly! They were part of the Indian national cricket team for almost 12 years. After Test series defeat at home against South Africa, Tendulkar resigned, and Sourav Ganguly took over as captain in 2000.

- legendary opening batting pair → (Coreference-partOf) → Sachin Tendulkar → (Anaphora-C) → They → (Coreference-RpartOf) → Indian National Cricket Team → (Coreference-partOf) → Tendulkar.
- legendary opening batting pair → (Coreference-partOf) → Sourav Ganguly → (Anaphora-C) → They → (Coreference-RpartOf) → Indian National Cricket Team → (Coreference-partOf) → Sourav Ganguly → (Coreference-noun-noun) → captain.
- blue colored texts are coreference mentions in above paragraph.

Coreference Types

- Anaphor
  - Concrete Anaphor
  - Abstract/Event Anaphor
  - Temporal Anaphor
- Coreference
  - Nominal Reference
  - Verbal Reference
  - Verb-nominal/Noun-Verbal reference

Dataset

The following statistics are from Hindi Dependency Treebank:

-	Docs	Sens	Tkns	Mentions
HDTB	1100	15476	185000	21354

Following table shows the coreference relation types distribution.

Type	Occurrences
Anaphora-C	2205
Anaphora-E	156
Anaphora-T	50
Anaphora-RC	72
Anaphora-RE	89
Anaphora-RT	87
Coreference-partOf	659
Coreference-Inferred	107
Coreference-Function-Value	67
Coreference-Identity (strong)	4525
Coreference-NounComplement	733
Coreference-Abbreviation	276
Coreference-Noun-Noun	943
Coreference-Noun-Verb	146

Coreference Resolution (system)

Three step process

1. Mention and mention head identifier
2. Anaphora/pronoun resolution
   1. Pronominal reference type identification
   2. Concrete anaphora resolution
   3. Event anaphora resolution
3. Nominal co-reference resolution
   1. Dependency relation match
   2. String match
   3. Abbreviation match
   4. Coref-Dic match
   5. String match with wordnet
   6. Word2vec and glove match
4. Relation type identification between continuous mentions of same chain

Features for pronominal type identification

- Pronoun,
- Pronoun's lexical distance from last verb,
- pronoun's lexical distance from upcoming verb,
- Pronoun's lexical position in sentence,
- Pronoun root,
- Pronoun's sentence position in discourse,
- Pronoun's chunk position in sentence,
- Pronoun's gender-number-person,
- Pronoun's next word and its category,
- Voice and sentence type of next and previous sentence,
- Pronoun's lexical distance from last occurred named entity,
- Lexical item count of pronouns chunk.

* These features are used in both rule base and learning base modules.

Features for Nominal reference resolution

- **1.** mention1 head, **2.** mention2 head, **3.** mention1 GNP, **4.** mention2 GNP, **5.** current mention1 is first mention of chain or not, **6.** current mention1 is second mention of chain or not, **7.** previous relation type between last mention pairs, **8.** mention1 to mention2 lexical distance, **9.** mention1 to mention2 chunk/phrase distance, **10.** mention1 to mention2 sentence distance,

* These features are used in both rule base and learning base modules.

Results

Sieve	Recall	Precision	F-Score
Mention idn.	94.83	85.85	90.11
Head idn.	-	-	93.89 (Acc)

Sieve	chains
Mention identification and detection	137(137)
Mention head identification	137(137)
Sieve 1 : Dependency relation match	127 (117)
Sieve 2 : String match	104 (107)
Sieve 3 : Abbreviation match	103 (107)
Sieve 4 : Coref-Dic String match	92 (67)
Sieve 5 : String match with WordNet	92 (67)
Sieve 6 : Word2vec and Glove match	87 (67)

Overall results for coreference resolution are recall 63.7, precision 79.53 and f-score 70.

Conclusion

In future, we will try to use this system in applications like question answering and discourse machine translation.

Authors

- Vandan Mujadia , LTRC (IIIT-H)
- Prof. Dipti M. Sharma , LTRC (IIIT-H)