Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution

Chen Chen and Vincent Ng

Human Language Technology Research Institute

The University of Texas at Dallas

Our Participation

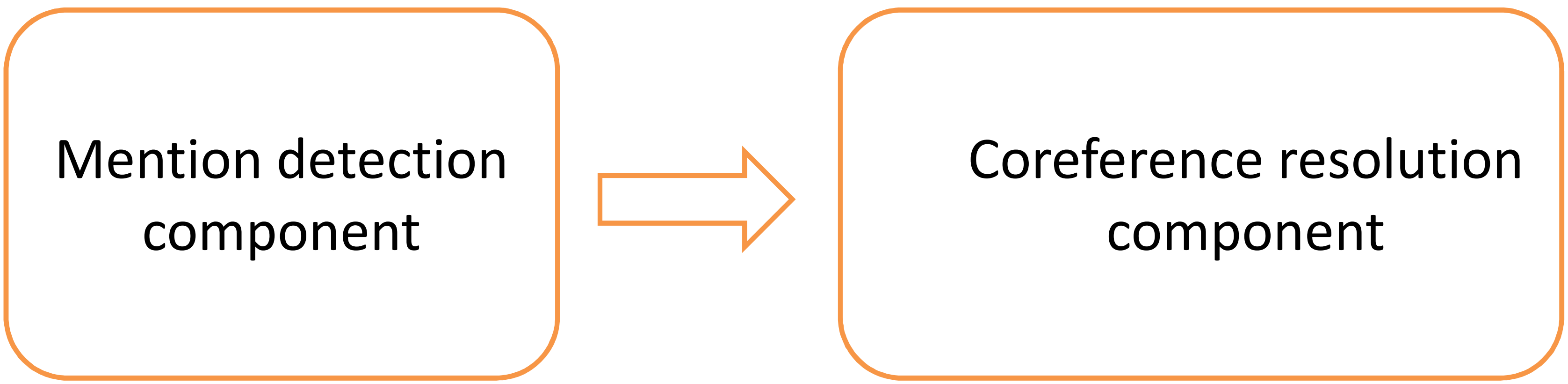
- Participated in 4 tracks
  - English (closed)
  - Chinese (closed)
  - Chinese (open)
  - Arabic (closed)

Major Results

- Official score on test set: 56.35
  - Ranked 3rd overall
- Ranked 1st in Chinese open and closed tracks

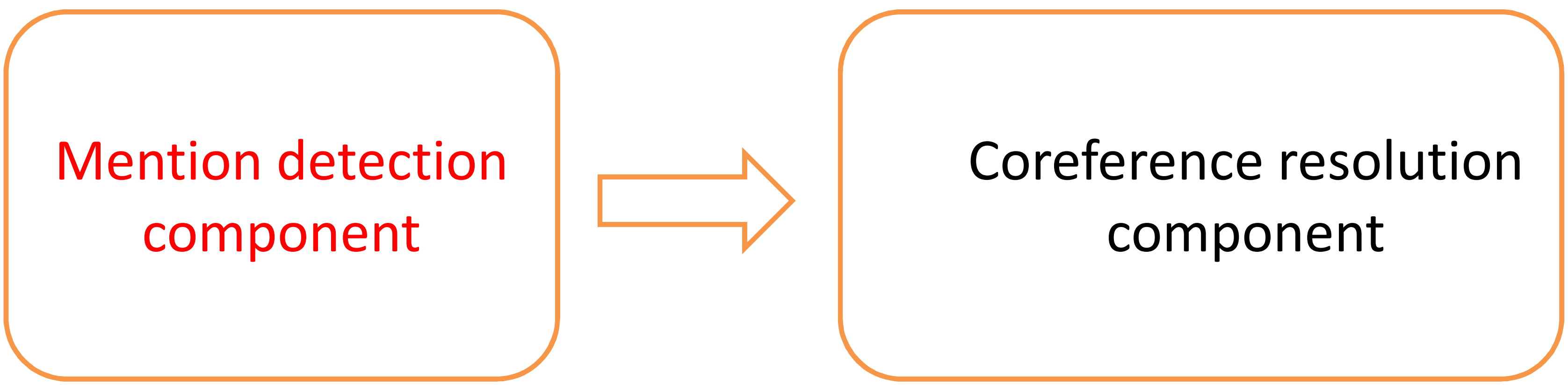
System Architecture

- A pipeline architecture



System Architecture

- A pipeline architecture



Mention Detection Component

- A hybrid approach
  - Combines rules and machine learning
- A three-step approach
  1. **Extraction** (improves recall)
     - Use parse trees and named entities to extract as many mentions as possible
  2. **Heuristic-based Pruning** (improves precision)
     - Heuristically prune erroneous mentions
  3. **Learning-based Pruning** (further improves precision)
     - Use training data to guide pruning

Mention Detection Component

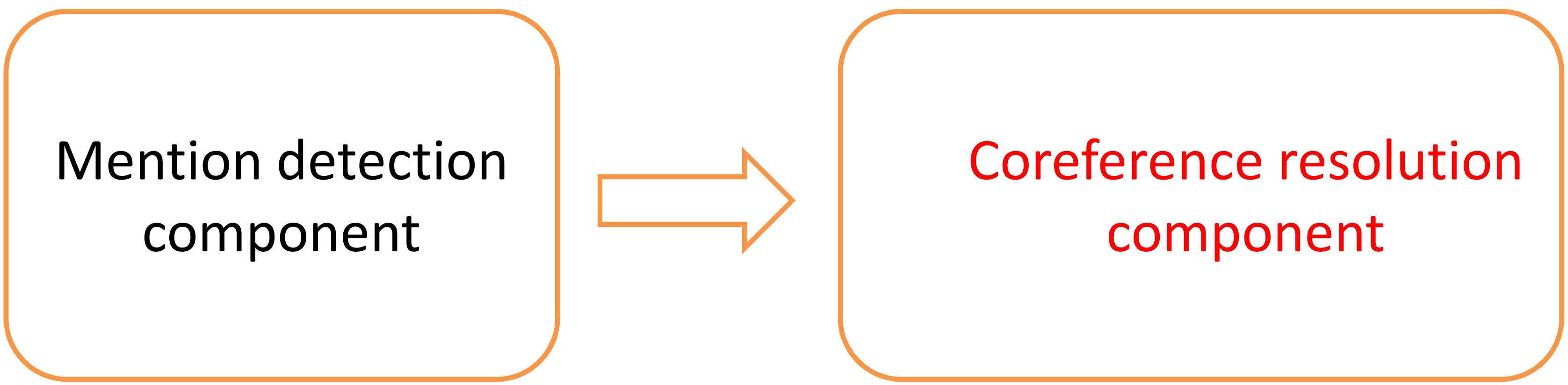
- A hybrid approach
  - Combines rules and machine learning
- A three-step approach
  1. **Extraction** (improves recall)
     - Use parse trees and named entities to extract as many mentions as possible
  2. **Heuristic-based Pruning** (improves precision)
     - Heuristically prune erroneous mentions
  3. **Learning-based Pruning** (further improves precision)
     - Use labeled data to guide pruning

Learning-Based Pruning

- Observation
  - If an NP is never annotated as a mention in the training data, it is probably not a mention
    - e.g., “no problem”, “the same”
- Learning-based pruning
  - Prune an extracted mention if the likelihood that its head is a mention (according to the training data) is less than the **Pruning Threshold**
    - a threshold to be tuned based on development data

System Architecture

- A pipeline architecture



Coreference Resolution Component

- Hybrid approach
  - Combines rule-based and learning-based methods
    1. Build a rule-based resolver
    2. Parameterize the resolver
    3. Learn the parameters
       - by leveraging training data

Coreference Resolution Component

- Hybrid approach
  - Combines rule-based and learning-based methods
    1. Build a rule-based resolver
    2. Parameterize the resolver
    3. Learn the parameters
       - by leveraging training data

Step 1: Build a Rule-Based Resolver

- use Stanford's multi-pass sieve approach
  - contains Stanford's sieves and our new sieves

Sieves for Chinese

- Chinese Head Match
- Discourse Processing
- Exact String Match
- Precise Constructs
- Strict Head Match A-*C*
- Proper Head Match
- Pronouns
- Lexical Pair Sieve

Chinese Head Match Sieve

- Applicable to only the newswire documents
  - owing to the way these documents are annotated
- Posits two mentions as coreferent if they have the same head and one is embedded within the other
- Exception: coordinated NP
  - *查尔斯和戴安娜[Charles and Diana] and 戴安娜 [Diana]*

Precise Constructs Sieve

- a Stanford sieve that posts two NPs as coreferent if one is an acronym or abbreviation of the other, or if they are appositives.

- We augment this sieve with additional rules to handle abbreviations in Chinese
  - **Abbreviation of foreign person names:**
  萨达姆·侯赛因[Saddam Hussein] and 萨达姆[Saddam]
  - **Abbreviation of Chinese person names:**
  陈总统[Chen President] and 陈水扁总统[Chen Shui-bian President].
  - **Abbreviation of country names**
  多国[Do country] and 多米尼加[Dominica]

Pronouns Sieve

- a Stanford sieve for resolving pronouns based on gender, number, and animacy agreement

- But … these three grammatical attribute values were not provided by the organizers for Chinese
  - We learned these values from the training data

How to learn these attribute values?

- 3 steps
  - Employ simple heuristics to extract attribute values for easy-to-handle mentions
    - *E.g., 她[she]* (Female, Single and Animate)
  - If a mention in a coreference chain has these attribute values extracted, we propagate such information to all mentions in the same chain
  - Based on these automatically extracted attribute values, we create six word lists: (1) animate words, (2) inanimate words, (3) female words, (4) male words, (5) singular words, and (6) plural words.

Lexical Pair Sieve

- Motivation
  - String/head match used in the Stanford sieves are not accurate indicators of coreference/non-coreference
  - Two mentions with same head may not be coreferent
    - E.g., *"别人[other people]" and "别人[other people]"* .
  - Two mentions with different heads may be coreferent
    - E.g., *"大陆[mainland]" and "中国[China]".*

Lexical Pair Sieve

- posits two mentions as coreferent if the probability they are coreferent (according to training data) >= S-high

- disallows two mentions to be coreferent if the probability they are coreferent <= S-low

- S-high, S-low tuned based on development data

Sieves for English

- Stanford sieves + Lexical Pair sieve

Sieves for Arabic

- Exact String Match sieve + Lexical Pair sieve
- Adding more sieves deteriorates performance

Coreference Resolution Component

- Hybrid approach
  - Combines rule-based and learning-based methods
    1. Build a rule-based resolver
    2. Parameterize the resolver
    3. Learn the parameters

Step 2: Parameterize the Resolver

- Two sets of tunable parameters
  - Lexical probability thresholds
    - E.g., S-low, S-high, Pruning Threshold
  - Rule relaxation parameters
    - each condition of a coreference rule in each sieve is associated with a parameter to control whether the condition should be removed or not
      - Can potentially simplify a rule

Coreference Resolution Component

- Hybrid approach
  - Combines rule-based and learning-based methods
    1. Build a rule-based resolver
    2. Parameterize the resolver
    3. Learn the parameters

Step 3: Learn the Parameters

- The two types of parameters are learned jointly to optimize the desired evaluation measure (average of MUC, CEAF, and B^3) on development data

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1

- Performance drops when either set of parameters is removed from the system

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1
Open	Full	72.9	65.3	74.8	50.7	**63.6**
Open	-Rule relaxation parameters	72.8	65.1	74.5	50.4	63.3
Open	-Lexical probability thresholds	72.7	65.0	74.5	50.4	63.3
Open	-Rule relaxation parameters -Lexical probability thresholds	72.4	64.6	74.3	50.1	63.0

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1
Open	Full	72.9	65.3	74.8	50.7	**63.6**
Open	-Rule relaxation parameters	72.8	65.1	74.5	50.4	63.3
Open	-Lexical probability thresholds	72.7	65.0	74.5	50.4	63.3
Open	-Rule relaxation parameters -Lexical probability thresholds	72.4	64.6	74.3	50.1	63.0

- Open track: resolver employs named entity information
  - Consistent improvement in performance
  - Both sets of parameters are crucial to performance

Chinese Development Set F-Scores

Track	System	MD	MUC	BCUBED	CEAF	AVG
Closed	Full	72.4	64.1	74.1	50.5	**62.9**
Closed	-Rule relaxation parameters	71.9	64.2	74.0	49.9	62.6
Closed	-Lexical probability thresholds	71.9	63.5	73.8	50.0	62.4
Closed	-Rule relaxation parameters -Lexical probability thresholds	71.5	63.3	73.6	49.5	62.1
Open	Full	72.9	65.3	74.8	50.7	**63.6**
Open	-Rule relaxation parameters	72.8	65.1	74.5	50.4	63.3
Open	-Lexical probability thresholds	72.7	65.0	74.5	50.4	63.3
Open	-Rule relaxation parameters -Lexical probability thresholds	72.4	64.6	74.3	50.1	63.0

- Similar trends observed for English and Arabic

Official Test Set F-Scores

Track	Track	MD	MUC	BCUBED	CEAF	AVG
English	Closed	73.8	63.7	69.0	46.4	59.7
Chinese	Closed	71.6	62.2	73.6	51.0	62.2
Chinese	Open	72.4	64.7	74.6	51.3	63.5
Arabic	Closed	59.8	39.0	61.5	40.8	47.1

Official Test Set F-Scores

Track	Track	MD	MUC	BCUBED	CEAF	AVG
English	Closed	73.8	63.7	69.0	46.4	59.7
Chinese	Closed	71.6	62.2	73.6	51.0	62.2
Chinese	Open	72.4	64.7	74.6	51.3	63.5
Arabic	Closed	59.8	39.0	61.5	40.8	47.1

- Best result in Chinese closed and open tracks
  - NE information useful for Chinese coreference
- Results for Arabic are fairly poor
  - Due to lack of linguistic expertise

Conclusion

- Proposed a hybrid rule-based and learning-based approach to coreference resolution
- Showed that the learning-based multi-pass sieve approach can work well for Chinese
- Feature engineering plays an important role in performance
  - But this requires language specific knowledge