Chinese Overt Pronoun Resolution: A Bilingual Approach

Chen Chen and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

AAAI-2014

Task: Chinese Overt Pronoun Resolution (PR)

Find an antecedent for each anaphoric overt pronoun in a Chinese text

An Illustrative Example

玛丽告诉约翰她非常喜欢他.

Mary told John that she liked him a lot.

Resolve pronouns 她**(she)** and 他**(him)** to their antecedents, which are 玛丽**(Mary)** and 约翰**(John)** respectively.

Why is it more Challenging than English PR?

- Less coreference-annotated data available for training resolvers in Chinese than in English
- Lack of publicly-available linguistic resources in Chinese that are essential for overt PR, such as Gender and Number wordlists

Goal: Improve Chinese PR

Address the two challenges above by exploiting
- English coreference-annotated data, and
- English Gender and Number wordlists

in addition to Chinese coreference-annotated data

How?

- Idea 1: Feature augmentation
  - Machine-translate Chinese text to English text
  - Align the Chinese and English mentions
  - Train a Chinese pronoun resolver on Chinese data, where instances are represented by features derived from Chinese mentions and those derived from the mapped English mentions
  - **Pros**: use Chinese training data; use English wordlists
  - **Cons**: does not use English training data
- Idea 2: Annotation projection
  - Train an English pronoun resolver on English data
  - Apply resolver on English text machine-translated from Chinese
  - **Pros:** use English training data; use English wordlists
  - **Cons:** does not use Chinese training data
- Idea 3: Our bilingual approach
  - Combine ideas 1 and 2 via an ensemble approach

Related Work

- All existing approaches to Chinese overt PR or coreference resolution are monolingual, training models on either
  - **Chinese** data (e.g., Luo and Zitouni (2005); Wang and Ngai (2006); Kong and Zhou (2012); Kong and Ng (2013)); or
  - **English** data (by adopting Idea 2) (Rahman and Ng, 2012),

  but none of them exploits resources in both languages

Bilingual Approach: Implementation Details

Document preprocessing

Step 1: Machine-translate each training and test document from Chinese to English using Google Translate

玛丽告诉约翰她非常喜欢他.

Mary told John that she liked him a lot.

Step 2: Align the words in each pair of sentences using BerkeleyAligner

Step 3: Align Chinese mentions to English mentions heuristically

[玛丽]告诉[约翰][她]非常喜欢[他].

[Mary] told [John] that [she] liked [him] a lot.

Classifier training (3 classifiers, all mention-pair models)

- English pronoun resolver (***PR^E***)
  - Trained on English training data
  - Training instances created from English anaphoric pronouns
  - Employs the English features from Björkelund and Farkas (2012)
- Chinese pronoun resolver (***PR^C***)
  - Trained on Chinese training data
  - Training instances created from Chinese anaphoric pronouns
  - Employs the Chinese features from Björkelund and Farkas (2012)
- Mixed pronoun resolver (***PR^M***)
  - Trained on Chinese training data and translated English data
  - Training instances created from the subset of Chinese anaphoric pronouns that have been aligned to some English pronouns
  - Employs the features used in both ***PR^E*** and ***PR^C***

Resolution methods

- Method 1
  - Resolve pronoun m_k to the closest preceding mention m_j whose coreference probability with m_k according to ***PR^E*** is at least 0.5
  - If m_k is not aligned to any English pronoun or ***PR^E*** does not resolve m_k, apply ***PR^C*** to resolve m_k
- Method 2
  - Same as method 1 except that ***PR^E*** is replaced with ***PR^M***
- Method 3
  - Same as method 1, except that the coreference probability between m_k and m_j is computed as the unweighted average of the probabilities returned by ***PR^E***, ***PR^M*** and ***PR^C*** (which we will refer to as P^E, P^M, and P^C respectively)
- Method 4
  - Resolve m_k to the closest preceding mention if at least one of four conditions is satisfied: (1) $P^C>t_C$; (2) $P^M>t_M$; (3) $P^E>t_E$ and $P^{norm}\geq 0.5$

$$P_{jk}^{norm} = \frac{P_{jk}^C + P_{jk}^E w_e(P_a)^{w_{ae}} + P_{jk}^M w_m(P_a)^{w_{am}}}{1 + w_e(P_a)^{w_{ae}} + w_m(P_a)^{w_{am}}}.$$

Note: P_a is the mention alignment probability, and t_C, t_M, t_E, w_e, w_{ae}, w_m, w_{am} are parameters tuned using the development set

Evaluation

Corpus: CoNLL-2012 shared task data

- Training set: 1,391 Chinese docs (750K words); 1,940 English docs (1.3M words)
- Development set: 172 Chinese docs (110K words)
- Test set: 166 Chinese docs (90K words)

Baseline systems

- Monolingual approach
  - Supervised mention-pair model trained only on Chinese data
- Best Chinese coreference system in the CoNLL-2012 shared task
  - Hybrid model combining rules and machine learning (Chen and Ng, 2012)
- Rahman and Ng's (2012) approach
  - Annotation projection approach (method 1 without using ***PR^C*** as backoff)

Evaluation metrics

- Recall (**R**), precision (**P**), and F-score (**F**) on resolving anaphoric pronouns
- Accuracies: $\mathbf{A^a}$ is the percentage of anaphoric pronouns correctly resolved; $\mathbf{A^{na}}$ is the percentage of non-anaphoric pronouns not resolved; $\mathbf{A^o}$ is overall accuracy

Results on CoNLL-2012 shared task test set

Resolution Method	R	P	F	A^a	A^{na}	A^o
Monolingual Approach (Closest-first)	71.7	65.3	68.4	71.7	59.3	67.4
Monolingual Approach (Best-first)	72.0	65.6	68.7	72.0	59.3	67.6
Best Shared Task System	63.8	67.5	65.6	63.8	**76.7**	68.2
Rahman and Ng's (2012) Approach	64.3	65.2	64.7	64.3	68.5	65.8
Method 1	65.6	64.4	65.0	65.6	66.0	65.8
Method 2	**73.0**	65.1	68.8	**73.0**	56.7	67.4
Method 3	71.5	70.5	71.0	71.5	67.6	70.2
Method 4	71.1	**71.5**	**71.3**	71.1	70.4	**70.8**

Methods 3 and 4 significantly outperform the baselines w.r.t. both F-score and accuracy.

Impact of Machine Translation Quality

5-fold cross-validation results on a 400-document parallel corpus

	Machine Translation (MT)			Human Translation (HT)		
Resolution Method	R	P	F	R	P	F
Monolingual Approach (Closest-first)	63.0	62.7	62.8	63.0	62.7	62.8
Monolingual Approach (Best-first)	62.3	62.0	62.2	62.3	62.0	62.2
Best Shared Task System	55.2	65.8	60.1	55.2	65.8	60.1
Rahman and Ng's (2012) Approach	54.7	58.1	56.4	46.1	59.9	52.1
Method 1	55.6	57.4	56.5	56.3	59.8	58.0
Method 2	**65.6**	59.8	62.5	**65.7**	61.7	63.7
Method 3	61.7	**66.9**	64.2	63.6	66.7	65.1
Method 4	63.8	65.3	**64.5**	64.5	**67.0**	**65.7**

When MT is replaced with HT, the F-scores of all four methods increase significantly by 0.9-1.5%, but their relative performance does not change.