**Task:** Chinese Overt Pronoun Resolution (PR)

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

**An Illustrative Example**

Mary told John that she liked him a lot.

Resolve pronouns **she** and **him** to their antecedents, which are **Mary** and **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.

Mary 告诉 John 她非常喜欢他.

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

Mary told John that she liked him a lot.

**Step 3:** Align Chinese mentions to English mentions heuristically.

[Mary 告诉 John 她]非常喜欢[他].

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

- **English pronoun resolver (PR)**
  - 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)**
  - 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)**
  - 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 and PR

**Resolution methods**

- **Method 1**
  - Resolve pronoun \(m_p\) to the closest preceding mention \(m_h\) whose co-reference probability \(PR\) is at least 0.5.
  - If \(m_p\) is not aligned to any English pronoun or \(PR\) does not resolve \(m_p\), apply \(PR\) to resolve \(m_h\).
- **Method 2**
  - Same as method 1 except that \(PR\) is replaced with \(PR\).
- **Method 3**
  - Same as method 1 except that the co-reference probability between \(m_h\) and \(m_p\) is computed as the unweighted average of the probabilities returned by \(PR\), \(PR\), and \(PR\) (which we will refer to as \(PR\), \(PR\), and \(PR\), respectively).
- **Method 4**
  - Resolve \(m_h\) to the closest preceding mention \(m_h\) if at least one of four conditions is satisfied: (1) \(PR > PR\); (2) \(PR > PR\); (3) \(PR > PR\); and 

\[
Pr_{new} = \frac{P_{PR} + P_{PR} w_s(P_s)^{\gamma} + P_{PR} w_m(P_m)^{\gamma}}{1 + w_s(P_s)^{\gamma} + w_m(P_m)^{\gamma}}
\]

**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 \(PR\) as backoff)

**Evaluation metrics**

- **Recall** (R), **Precision** (P), and **F-score** (F) on resolving anaphoric pronouns
- **Accuracies:** A is the percentage of anaphoric pronouns correctly resolved; A is the percentage of non-anaphoric pronouns not resolved; A is overall accuracy.

**Results on CoNLL-2012 shared task test set**

<table>
<thead>
<tr>
<th>Resolution Method</th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual Approach</td>
<td>71.7</td>
<td>65.3</td>
<td>68.4</td>
</tr>
<tr>
<td>Monolingual Approach (Best-first)</td>
<td>72.0</td>
<td>65.6</td>
<td>68.7</td>
</tr>
<tr>
<td>Best Shared Task System</td>
<td>63.8</td>
<td>67.5</td>
<td>65.6</td>
</tr>
<tr>
<td>Rahman and Ng’s (2012) Approach</td>
<td>64.3</td>
<td>65.2</td>
<td>64.7</td>
</tr>
<tr>
<td><strong>Method 1</strong></td>
<td>65.6</td>
<td>64.4</td>
<td>65.0</td>
</tr>
<tr>
<td><strong>Method 2</strong></td>
<td>73.0</td>
<td>65.1</td>
<td>68.8</td>
</tr>
<tr>
<td><strong>Method 3</strong></td>
<td>71.5</td>
<td>70.5</td>
<td>71.0</td>
</tr>
<tr>
<td><strong>Method 4</strong></td>
<td>71.1</td>
<td>71.5</td>
<td>71.3</td>
</tr>
</tbody>
</table>

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

**Impact of Machine Translation Quality**

**S-fold cross-validation results on a 400-document parallel corpus**

<table>
<thead>
<tr>
<th>Resolution Method</th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation (MT)</td>
<td>63.0</td>
<td>62.7</td>
<td>62.8</td>
</tr>
<tr>
<td><strong>Human Translation (HT)</strong></td>
<td>63.0</td>
<td>62.7</td>
<td>62.8</td>
</tr>
<tr>
<td>Monolingual Approach (Closest-first)</td>
<td>62.3</td>
<td>62.0</td>
<td>62.2</td>
</tr>
<tr>
<td>Monolingual Approach (Best-first)</td>
<td>55.2</td>
<td>65.8</td>
<td>60.1</td>
</tr>
<tr>
<td>Best Shared Task System</td>
<td>54.7</td>
<td>58.1</td>
<td>56.4</td>
</tr>
<tr>
<td>Rahman and Ng’s (2012) Approach</td>
<td>55.6</td>
<td>57.4</td>
<td>56.5</td>
</tr>
<tr>
<td><strong>Method 1</strong></td>
<td>65.6</td>
<td>69.8</td>
<td>62.5</td>
</tr>
<tr>
<td><strong>Method 2</strong></td>
<td>66.5</td>
<td>69.8</td>
<td>62.5</td>
</tr>
<tr>
<td><strong>Method 3</strong></td>
<td>67.1</td>
<td>66.9</td>
<td>64.2</td>
</tr>
<tr>
<td><strong>Method 4</strong></td>
<td>63.8</td>
<td>65.3</td>
<td>64.5</td>
</tr>
</tbody>
</table>

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.