Generating and Exploiting Large-scale Pseudo Training Data for Zero Pronoun Resolution

Ting Liu†, Yiming Cui‡, Qingyu Yin†, Weinan Zhang†, Shijin Wang‡ and Guoping Hu‡

†Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China
‡iFLYTEK Research, Beijing, China
Zero Pronoun (ZP)

小明 吃了 一个 苹果， <ZP> 非常 甜

Xiaoming eats an apple, it is very sweet
Zero Pronoun Resolution (ZPR)

小明 吃 了 一个 苹果，<ZP> 非常 甜

Xiaoming eats an apple, it is very sweet
Challenges of ZPR

- No overt pronoun for indication
  - No information for the positions of ZPs
  - No type/surface information of ZPs

- Feature engineering

<table>
<thead>
<tr>
<th>Syntactic features (13)</th>
<th>Other features (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>whether $z$ is the first gap in an IP clause; whether $z$ is the first gap in a subject-less IP clause, and if so, POS($w_1$); whether POS($w_1$) is NT; whether $w_1$ is a verb that appears in a NP or VP, whether $P_1$ is a NP node; whether $P_2$ is a VP node, the phrasal label of the parent of the node containing POS($w_1$); whether $V$ has a NP, VP or CP ancestor; whether $C$ is a VP node; whether there is a VP node whose parent is an IP node in the path from $w_1$ to $C$.</td>
<td>whether $z$ is the first gap in a sentence; whether $z$ is in the headline of the text, the type of the clause in which $z$ appears; the grammatical role of $z$ (SUBJECT, OBJECT, or OTHER); whether $w_{-1}$ is a punctuation; whether $w_{-1}$ is a comma.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Syntactic features (12)</th>
<th>Distance features (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>whether $e$ has an ancestor NP, and if so, whether this NP is a descendent of $e$’s lowest ancestor IP; whether $e$ has an ancestor VP, and if so, whether this VP is a descendent of $e$’s lowest ancestor IP; whether $e$ has an ancestor CP; the grammatical role of $e$ (SUBJECT, OBJECT, or OTHER); the clause type in which $e$ appears, whether $e$ is an adverbial NP, a temporal NP, a pronoun or a named entity.</td>
<td>the sentence distance between $e$ and $z$; the segment distance between $e$ and $z$, where segments are separated by punctuations; whether $e$ is the closest NP to $z$; whether $e$ and $z$ are siblings in the associated parse tree.</td>
</tr>
</tbody>
</table>

19 hand-crafted features for ZP

18 hand-crafted features for antecedent

Solutions

• No overt pronoun for indication
  • Considering all possible positions for ZPs identification
  • Classifying ZPs to Anaphoric ZPs (AZP) and Non-AZPs
  • Modelling the semantics of ZPs and antecedents

• Feature engineering
  • Automatically learning to represent features
  • Deep learning approaches for the modeling
  • More labeled data for training

Most existing work

This paper
How to Obtain Large-scale Training Data?

• Manual Annotation
  • Labor consuming
  • Hard to say “large-scale”

• Automatic Generation
  • Easy to obtain
  • Large-scale
  • Pseudo training data
What is Actual Training Data?

- Sample Training Data in OntoNotes 5.0
  - Single-word (In Chinese) antecedent
  
  CN: [警方] 怀疑 这是一起 黑枪 案件，zp₁ 将 枪械 交 送 市里 zp₂ 以 清理 案情。
  EN: [The police] suspected that this is a criminal case about illegal guns, zp₁ brought the guns to the city zp₂ to deal with the case.

- Multi-word antecedent

CN: 这次 [近 50 年 来 印度 发生 的 最 强烈 地震] 震级 强，zp 波 及 范围 广，印 度 邻国 如 尼泊尔 也 受到 了 影响。
EN: [The earthquake that is the strongest one occurs in India within recent 50 years] has a high-magnitude, zp influences a large range of areas, and the neighboring country of India like Nepal is also affected.
How to Generate Pseudo Training Data?

• Collecting large-scale news documents, which is relevant (or homogenous in some sense) to the OntoNotes 5.0 data.

• Given a document $D$, a word is randomly selected as an answer $A$ if
  • It is either a noun or pronoun
  • It should appear at least twice in the document

• The sentence contains $A$ is defined as a query $Q$, in which the answer $A$ is replaced by a specific symbol “<blank>”
welcome both of you to the studio to participate in our program,

it happened that i was going to have lunch with a friend at noon.

after that, i received an sms from 1860.

uh-huh, it was by sms.

uh-huh, that means, er, you knew about the accident through the source of radio station.

although we live in the west instead of the east part, and it did not affect us that much,

but i think it is very useful to inform people using sms.

some car owners said that <blank> was very good.

has owner said, this <blank> is very good.
Zero Pronoun Resolution (ZPR)

• A pseudo training sample can be represented as

\[ < D, Q, A > \]

• Zero pronoun resolution task is thus defined as

\[ P(A|D, Q) \]
Attention-based NN Model for ZPR

Two-step Training

- Single-word
  - Matching the single word
- Multi-word Antecedent
  - Matching the head word

Pseudo Data Pre-training

General Training

Actual Data Fine-tuning

Domain Training
Experimental Data

• OntoNotes Release 5.0 from CoNLL-2012
  • Broadcast News (BN), Newswires (NW), Broadcast Conversations (BC), Telephone Conversations (TC), Web Blogs (WB), Magazines (MZ)

<table>
<thead>
<tr>
<th></th>
<th>Sentences #</th>
<th>Query #</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Train</td>
<td>18.47M</td>
<td>1.81M</td>
</tr>
<tr>
<td>Domain Train</td>
<td>122.8K</td>
<td>9.4K</td>
</tr>
<tr>
<td>Validation</td>
<td>11,191</td>
<td>2,667</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Docs</th>
<th>Sentences</th>
<th>Words</th>
<th>AZPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>172</td>
<td>6,083</td>
<td>110K</td>
<td>1,713</td>
</tr>
</tbody>
</table>
Overall Performance

• F-score

<table>
<thead>
<tr>
<th></th>
<th>NW (84)</th>
<th>MZ (162)</th>
<th>WB (284)</th>
<th>BN (390)</th>
<th>BC (510)</th>
<th>TC (283)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kong and Zhou (2010)</td>
<td>34.5</td>
<td>32.7</td>
<td>45.4</td>
<td>51.0</td>
<td>43.5</td>
<td>48.4</td>
<td>44.9</td>
</tr>
<tr>
<td>Chen and Ng (2014)</td>
<td>38.1</td>
<td>31.0</td>
<td>50.4</td>
<td>45.9</td>
<td>53.8</td>
<td><strong>54.9</strong></td>
<td>48.7</td>
</tr>
<tr>
<td>Chen and Ng (2015)</td>
<td>46.4</td>
<td>39.0</td>
<td>51.8</td>
<td>53.8</td>
<td>49.4</td>
<td>52.7</td>
<td>50.2</td>
</tr>
<tr>
<td>Chen and Ng (2016)</td>
<td>48.8</td>
<td>41.5</td>
<td>56.3</td>
<td><strong>55.4</strong></td>
<td>50.8</td>
<td>53.1</td>
<td>52.2</td>
</tr>
<tr>
<td>Our Approach†</td>
<td><strong>59.2</strong></td>
<td><strong>51.3</strong></td>
<td><strong>60.5</strong></td>
<td>53.9</td>
<td><strong>55.5</strong></td>
<td>52.9</td>
<td><strong>55.3</strong></td>
</tr>
</tbody>
</table>
Effect of UNK Processing

(a) The weather today is not as pleasant as the weather of yesterday.
(b) The <unk> today is not as <unk> as the <unk> of yesterday.
(c) The <unk1> today is not as <unk2> as the <unk1> of yesterday.

<table>
<thead>
<tr>
<th></th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without UNK replacement</td>
<td>52.2</td>
</tr>
<tr>
<td>With UNK replacement</td>
<td>55.3</td>
</tr>
</tbody>
</table>
Effect of Domain Adaptation

<table>
<thead>
<tr>
<th></th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Pseudo Training Data</td>
<td>41.1</td>
</tr>
<tr>
<td>Only Task-Specific Data</td>
<td>44.2</td>
</tr>
<tr>
<td>Only Task-Specific Data + GloVe</td>
<td>50.9</td>
</tr>
<tr>
<td>Domain Adaptation</td>
<td>55.3</td>
</tr>
</tbody>
</table>
Error Analysis

• The impact of UNK words

CN: zp unk1 unk2 顶，将 unk3 和 unk4 的美景尽收眼底。
EN: zp successfully [climbed]_{unk1} the peak of [Taiping Mountain]_{unk2} to have a panoramic view of the beauty of [Hong Kong Island]_{unk3} and [Victoria Harbour]_{unk4}.

• Long distance between ZPs and antecedents

CN: [我] 帮不了那个人...（多于30个词）...那天结束后，zp 回到家中。
EN: [I] can’t help that guy ... (more than 30 words) ... After that day, zp return home.
Conclusion

• Generating and exploiting pseudo training data for ZPR
  • Inspired by the cloze-style reading comprehension

• Two-step training of the ZPR model for the use of the large scale pseudo training data

• A new State-of-the-Art approach on Chinese ZPR task
Thanks!
Questions and Advices?