Generalized Abbreviation Prediction with Negative Full Forms & Application on Chinese Web Search

Xu Sun, Wenjie Li, Fanqi Meng, Houfeng Wang

Peking University
The Hong Kong Polytechnic University
Outline

- Introduction
- Method
- Experiments
- Conclusion
Abbreviation processing

- Short form to full form: Abbreviation Expansion
- Extracting short form & full form pairs from context: Abbreviation Recognition/Disambiguation
- **Full form to short form: Abbreviation prediction/generation**

*dynamic programming* $\rightarrow$ $DP$  
expanded form, full form, long form  
abbreviation, short form
Why abbreviation prediction is important?

- **Helpful for information retrieval**
  - Text may contain only abbreviations → searching for the full form is useless
  - In People’s Daily data, >70% articles contain only the abbreviation “欧盟”

<table>
<thead>
<tr>
<th>Full Form</th>
<th>Abbr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>欧洲经济与货币联盟 → 欧盟</td>
<td>European Economic and Monetary Union</td>
</tr>
</tbody>
</table>

- **Helpful for the abbreviation recognition task**
  - Abbreviation recognition → restricted abbreviation prediction [Xu Sun+ ACL 2009]
  - Better abbreviation prediction → better abbreviation recognition
Our Focus & Existing problem

- **Target**
  - accurate Chinese abbreviation prediction for real-world applications

- **Existing problem**
  - Existing techniques focus on positive full forms only!
  - Positive full form $\rightarrow$ which will surely have abbreviation
  - Why? $\rightarrow$ this lab setting makes the task easier
  - But.. real world applications contains negative full forms
Our Focus & Existing problem

- **Target**
  - accurate Chinese abbreviation prediction for real-world applications

- **Existing problem**
  - Existing techniques focus on positive full forms only!
  - Positive full form → will surely have abbreviation
  - Why? → this lab setting makes the task easier
  - But.. real world applications contains negative full forms

<table>
<thead>
<tr>
<th>Positive/Negative Full Forms</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>磷酸氢二钠</td>
<td>X</td>
</tr>
<tr>
<td>君主专制</td>
<td>X</td>
</tr>
<tr>
<td>珠穆朗玛峰</td>
<td>珠峰</td>
</tr>
<tr>
<td>天公不作美</td>
<td>X</td>
</tr>
<tr>
<td>中国社会科学院</td>
<td>中国社科院</td>
</tr>
<tr>
<td>新时期的总任务</td>
<td>X</td>
</tr>
<tr>
<td>自由民主党</td>
<td>自民党</td>
</tr>
<tr>
<td>车辆发动机</td>
<td>X</td>
</tr>
<tr>
<td>复员退伍军人安置办公室</td>
<td>复退办</td>
</tr>
<tr>
<td>丰尔尼雪夫斯基</td>
<td>X</td>
</tr>
<tr>
<td>持谨慎态度</td>
<td>X</td>
</tr>
<tr>
<td>土产日用品杂品公司</td>
<td>X</td>
</tr>
<tr>
<td>一叶蔽目不见泰山</td>
<td>X</td>
</tr>
<tr>
<td>打击黑势力但除恶势力</td>
<td>打黑扫恶</td>
</tr>
</tbody>
</table>
Generalized Abbreviation Prediction with Negative Full Forms

- Can effectively deal with mixed positive/negative full forms
- Thus can deal with real-world applications containing both positive & negative full forms
Outline

- Introduction
- Method
- Experiments
- Conclusion
Generalized Abbreviation Prediction

- **Preprocessing**
  - Word segmentation for full forms
  - Part-of-speech tagging for full forms

- **Unified system for generalized abbr. predict.**
  - Can do abbr. predict. in 1 single step
  - Casting abbr. predict. as a sequence labeling problem

Mount Everest

珠穆朗玛峰

Y N N N Y  →  珠峰
Generalized Abbreviation Prediction

- **Preprocessing**
  - Word segmentation for full forms
  - Part-of-speech tagging for full forms

- **Unified system for generalized abbr. predict.**
  - Can do abbr. predict. in 1 single step
  - Casting abbr. predict. as a sequence labeling problem → using the well-known CRF model
  - For negative full forms → 2 assumptions
    - **Assumption-1:** assuming its abbr. is nothing (null-string)
    - **Assumption-2:** assuming its abbr. is the full form itself
    - Which assumption is better?
Features & label-encoding

- **Features**
  - Character features
  - Character bi-gram
  - Numeral
  - Organization name suffix
  - Location name suffix
  - Word segmentation information
  - POS-tagging information

- **Label encoding with global information**
  - Chinese abbr. generation is highly dependent on global info
  - E.g., the number of characters of the generated abbreviations
  - Well-solved by the **GI method [Xu Sun+ ACL 2009]**, thus GI is adopted in this work as well
Training the system

- **Batch training**
  - Limited memory BFGS (LBFGS)
  - Too slow, need 300 passes for training

- **Existing online training methods**
  - Stochastic gradient descent (SGD)
  - Moderately fast, need 50 passes for training

- **Our method: feature-frequency-adaptive online training ADF** [Xu Sun+ ACL 2012], [Xu Sun+, Computational Linguistics, 2013, to appear]
  - General-purpose; Fast & accurate training
  - Finish the training in 10 passes
Applying generalized abbr. predict. tech for improving Chinese web search

Query expansion
- Full form + predicted abbreviation

Query alternation
- Full form → predicted abbreviation
Outline

- Introduction
- Method
- Experiments
- Conclusion
Experiments

- **Data**
  - We build a dataset containing 10,786 full forms collected from People’s Daily corpora & SIGHAN word segmentation corpora
  - 8,015 positive full forms w/ abbreviations annotated
  - 2,771 negative full forms

<table>
<thead>
<tr>
<th>Category</th>
<th>Portion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun Phrase</td>
<td>52.01%</td>
</tr>
<tr>
<td>Verb Phrase</td>
<td>13.72%</td>
</tr>
<tr>
<td>Organization Name</td>
<td>26.84%</td>
</tr>
<tr>
<td>Location Name</td>
<td>5.28%</td>
</tr>
<tr>
<td>Person Name</td>
<td>0.32%</td>
</tr>
<tr>
<td>Others</td>
<td>1.80%</td>
</tr>
</tbody>
</table>
Experiments on generalized abbr. predict.
- Randomly sampled 8,629 samples (80%) for training
- 2,157 (20%) for testing
- All-match accuracy (All-Acc)
  - #correct outputs (i.e., label strings) divided by #full forms in the test set
- Character accuracy (Char-Acc)

Experiments on web search
- the 2,157 testing samples as queries for web search
- Based on “title search” of “baidu.com”
- Precision, Recall, F-score
- Macro-averaging, micro-averaging
Results: generalized abbr. predict.

<table>
<thead>
<tr>
<th>Method</th>
<th>Discriminate Acc (%)</th>
<th>Overall All-Acc</th>
<th>Overall Char-Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic System</td>
<td>73.20</td>
<td>25.77</td>
<td>65.79</td>
</tr>
<tr>
<td>Unified-Assum.1 (Perc)</td>
<td>87.48</td>
<td>54.89</td>
<td>87.02</td>
</tr>
<tr>
<td>Unified-Assum.1 (MEMM)</td>
<td>86.97</td>
<td>50.16</td>
<td>85.92</td>
</tr>
<tr>
<td>Unified-Assum.1 (CRF-ADF)</td>
<td>87.80</td>
<td>56.69</td>
<td>87.20</td>
</tr>
<tr>
<td>Unified-Assum.1-GI (Perc)</td>
<td>91.93</td>
<td>75.42</td>
<td>90.23</td>
</tr>
<tr>
<td>Unified-Assum.1-GI (MEMM)</td>
<td>88.59</td>
<td>70.32</td>
<td>88.21</td>
</tr>
<tr>
<td>Unified-Assum.1-GI (CRF-ADF)</td>
<td>91.05</td>
<td><strong>79.46</strong></td>
<td><strong>91.61</strong></td>
</tr>
<tr>
<td>Unified-Assum.2 (Perc)</td>
<td>86.83</td>
<td>55.86</td>
<td>82.20</td>
</tr>
<tr>
<td>Unified-Assum.2 (MEMM)</td>
<td>87.52</td>
<td>56.18</td>
<td>82.27</td>
</tr>
<tr>
<td>Unified-Assum.2 (CRF-ADF)</td>
<td>87.11</td>
<td>56.97</td>
<td>82.54</td>
</tr>
<tr>
<td>Unified-Assum.2-GI (Perc)</td>
<td>90.35</td>
<td>71.85</td>
<td>88.04</td>
</tr>
<tr>
<td>Unified-Assum.2-GI (MEMM)</td>
<td>87.99</td>
<td>63.74</td>
<td>83.77</td>
</tr>
<tr>
<td>Unified-Assum.2-GI (CRF-ADF)</td>
<td>90.77</td>
<td>74.78</td>
<td>89.19</td>
</tr>
</tbody>
</table>

- Discriminate accuracy is high
- Heuristic system works poorly
- Global info (GI encoding) is helpful
- For negative full forms, Assumption-1 works better than Assumption-2
Why Assumption-1 works better than Assumption-2 for NFF?

- Probably because NFFs have no similar patterns with the real abbreviations
- #char in NFFs is very different compared with real abbr.
- NFFs contain formal word units vs. abbr. contain few word units
- Assumption-1 does not have such problems
Results: web search

- Prec: #correct-search-results divided by #total-search-results
- Rec: #correct-search-results divided by #total-existing-correct-search-results
  - #existing-correct-search-results ← estimated by #search-results of the full form + #search-results of the abbr.
- # is based on top10 returned pages (200 items)

<table>
<thead>
<tr>
<th>Method</th>
<th>Micro Prec</th>
<th>Micro Rec</th>
<th>Micro F1</th>
<th>Macro Prec</th>
<th>Macro Rec</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original query</td>
<td>48.51</td>
<td>18.14</td>
<td>26.41</td>
<td>47.84</td>
<td>28.76</td>
<td>35.92</td>
</tr>
<tr>
<td>Query alternation</td>
<td>47.73</td>
<td>54.04</td>
<td>50.69</td>
<td>62.84</td>
<td>61.12</td>
<td>61.97</td>
</tr>
<tr>
<td>Query expansion</td>
<td>47.93</td>
<td>72.18</td>
<td><strong>57.60</strong></td>
<td>53.70</td>
<td>82.02</td>
<td><strong>64.90</strong></td>
</tr>
</tbody>
</table>

1. Query expansion based on generalized abbr. predict. is significantly better than baseline
2. Major improvement is from the recall
3. Query alternation has lower recall than query expansion
4. However, query alternation is also better than baseline
## Some examples

<table>
<thead>
<tr>
<th>Original search: 游泳协会 (swimming association)</th>
<th>Relaxed abbreviation prediction based search: 泳协</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 成都市游泳运动协会2012年工作年会召开</td>
<td>• 戴利将变成“英国田亮”?活动过多已引起泳协不满</td>
</tr>
<tr>
<td>• 游泳协会起诉苯胺泄露企业,谁该反思</td>
<td>• 戴利母亲回击泳协主席警告:是你让你保住的饭碗!</td>
</tr>
<tr>
<td>• 八旬老人天天下河冬泳 六安游泳协会有个“泳魂”</td>
<td>• 戴利想当跳槽主持人遭警告 英泳协:别本末倒置</td>
</tr>
<tr>
<td>• 六安市游泳协会首届冬泳比赛开赛</td>
<td>• 国内游泳强队海外特训热浪汹涌 澳泳协施压也无用</td>
</tr>
<tr>
<td>• 池州市游泳运动协会正式成立</td>
<td>• 美的续签中国泳协,冠军助阵美的元春促销签名惠</td>
</tr>
</tbody>
</table>
Results: web search

<table>
<thead>
<tr>
<th>Method</th>
<th>Micro Prec</th>
<th>Micro Rec</th>
<th>Micro F1</th>
<th>Macro Prec</th>
<th>Macro Rec</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original query</td>
<td>48.51</td>
<td>18.14</td>
<td>26.41</td>
<td>47.84</td>
<td>28.76</td>
<td>35.92</td>
</tr>
<tr>
<td>Query alternation (gold-standard)</td>
<td>72.31</td>
<td>81.86</td>
<td>76.79</td>
<td>83.07</td>
<td>79.11</td>
<td>81.04</td>
</tr>
<tr>
<td>Query expansion (gold-standard)</td>
<td>66.40</td>
<td>100.00</td>
<td>79.81</td>
<td>65.56</td>
<td>100.00</td>
<td>79.19</td>
</tr>
</tbody>
</table>

- checking up-bound F-score of generalized abbr. predict. for web search
- up-bound is achieved by 100% correct system → gold-standard abbreviations are used
- up-bound of micro-F-score & macro-F-score is 79.81% & 81.04%

Thus, the web search quality of generalized abbr. predict. still has a large space to be improved, possibly via larger training dataset in the future
This work is dedicated on generalized abbreviation prediction & its application on improving web search.

Experiments demonstrate that the unified system based on global information outperforms the baselines.

Experiments also demonstrate that generalized abbreviation prediction can improve web search qualities.

Future work:
- try to improve the performance via collecting more training data or via semi-supervised learning.
Thank you!

Any question?