# 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

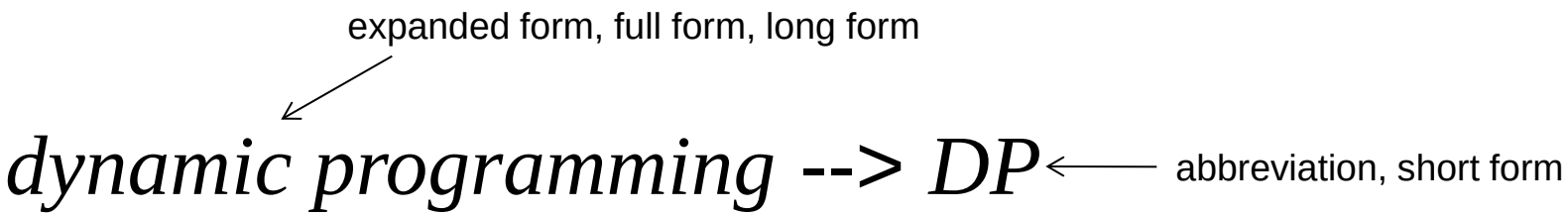

北京大学

# Outline

- Introduction
- Method
- Experiments
- Conclusion

# Introduction

- **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**



# 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"

| Full Form | Abbr. |
| --- | --- |
| 欧洲经济与货币联盟 ⟶ | 欧盟 |
| European Economic and Monetary Union | |

- **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 → which will surely have abbreviation
  - Why? → this lab setting makes the task easier
  - But.. real world applications contains negative full forms

# Our Focus & Existing problem

- **Target**
  - accurate Chinese abbr
    world applications
- **Existing problem**
  - Existing techniques fo
  - Positive full form → w
    abbreviation
  - Why? → this lab settin
  - But.. real world applic
    forms

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

# Proposal

- **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

# Method

## 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

# Method

## 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

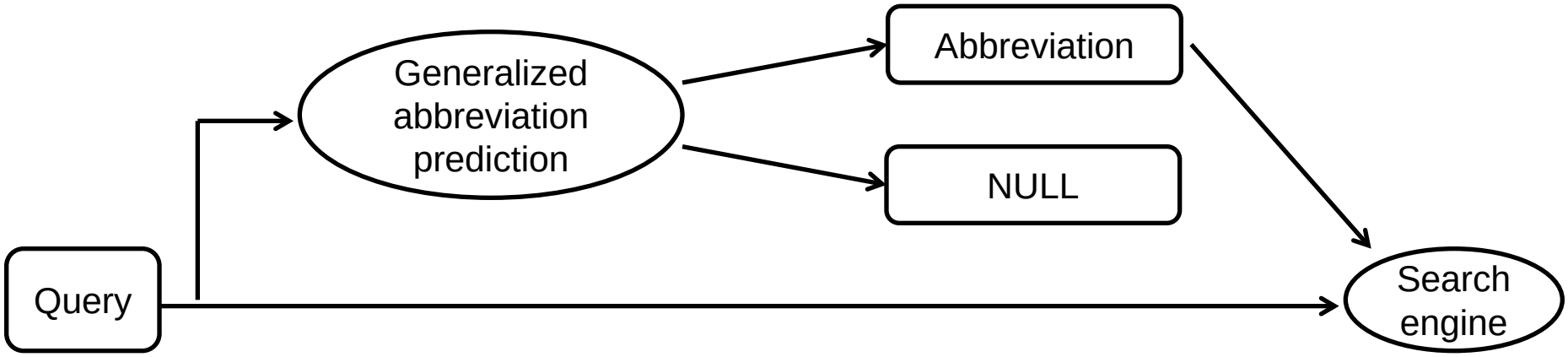
# Training the system

- **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

# Abbreviation Prediction for Web Search

- **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

| Category | Portion (%) |
|---|---|
| Noun Phrase | 52.01% |
| Verb Phrase | 13.72% |
| Organization Name | 26.84% |
| Location Name | 5.28% |
| Person Name | 0.32% |
| Others | 1.80% |

# Setting

- **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.

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

- **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**

# Discussion

- **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)**

| Method | Micro Prec | Micro Rec | Micro F1 | Macro Prec | Macro Rec | Macro F1 |
|---|---|---|---|---|---|---|
| Original query | 48.51 | 18.14 | 26.41 | 47.84 | 28.76 | 35.92 |
| Query alternation | 47.73 | 54.04 | 50.69 | 62.84 | 61.12 | 61.97 |
| Query expansion | 47.93 | 72.18 | **57.60** | 53.70 | 82.02 | **64.90** |

**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**

# Results: web search

## Some examples

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

# Results: web search

| Method | Micro Prec | Micro Rec | Micro F1 | Macro Prec | Macro Rec | Macro F1 |
|---|---|---|---|---|---|---|
| Original query | 48.51 | 18.14 | 26.41 | 47.84 | 28.76 | 35.92 |
| Query alternation (gold-standard) | 72.31 | 81.86 | 76.79 | 83.07 | 79.11 | **81.04** |
| Query expansion (gold-standard) | 66.40 | 100.00 | **79.81** | 65.56 | 100.00 | 79.19 |

- **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**

# Conclusions

- **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?**