The Design and Construction of A Chinese Collocation Bank

Ruifeng Xu, Qin Lu
* Dept. of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Introduction

- Design and construct a large scale and accurate collocation bank as a NLP resource for Chinese

Steps to build the collocation bank

- Classify Chinese collocations according to their different characteristics and features
- Design the guideline for collocation bank construction
- Construct collocation bank through manual annotation

Scale

- 3,643 headwords, 23,581 identical collocations in a 1-million word corpus
Collocation: Definition and Properties

A collocation is a recurrent and conventional expression containing two or more content word combinations that hold syntactic and/or semantic relations.

- Both uninterrupted collocation and interrupted collocation
- Both bi-gram collocation and n-gram collocation
- Properties
  - recurrent and habitual use: 历史包袱（historic baggage vs historic luggage），浓茶 (strong tea), but not 烈茶（powerful tea)
  - limited compositional: “裁减/v 员额/n” (reduce the posts)
  - limited substitutability and limited modifiability: 南/f 南/f 合作/vn (cooperation between south hemisphere countries)
  - domain-dependent: 专家/n 系统/n（expert systems）
## Classification of Chinese Collocations

- **Four types of collocations according to compositionality, substitutability, modifiability, and internal association**

<table>
<thead>
<tr>
<th></th>
<th>Type 0 Idiomatic Collocation</th>
<th>Type 1 Fixed Collocation</th>
<th>Type 2 Strong Collocation</th>
<th>Type 3 Loose Collocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compositional</td>
<td>No</td>
<td>Limited to yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Synonym substitutable</td>
<td>No</td>
<td>No</td>
<td>Very limited</td>
<td>Limited</td>
</tr>
<tr>
<td>Order alter</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Modifiable</td>
<td>No</td>
<td>No</td>
<td>Very Limited</td>
<td>Limited</td>
</tr>
<tr>
<td>Statistic significance</td>
<td>Not required</td>
<td>Not required</td>
<td>Required</td>
<td>Strongly Required</td>
</tr>
<tr>
<td>Examples</td>
<td>瘾木求鱼 (climb the tree to get fish – wrong method) 蓝牙 (Bluetooth)</td>
<td>n外交/ n 豁免权/ n(外交 immunity)</td>
<td>缔结/v 同盟/n 联盟/n (form alliances)</td>
<td>合法/v 收入/n 合法/v 收益/n (lawful income)</td>
</tr>
</tbody>
</table>
Guideline for Collocation Bank Construction

- The annotation follows headword-driven strategy.
- For a given headword, the annotation tasks include:
  - Identify its corresponding bi-gram collocations and n-gram collocations
  - Annotate and verify the occurrence of each collocation
    - Each occurrence must be manually examined because some are not collocations
  - For each bi-gram collocation, annotate its type and syntactic dependency relations
Annotation of Collocation Bank

- Corpus Data Preparation
  - People’s Daily Segmented Corpus (By Peking Univ.)

- Headword Set
  - 3,643 headwords selected from “The Dictionary of Modern Chinese Collocation”

- Each headword is annotated in three passes

- Syntactic Dependency Labels
  - **PZA** - Noun and its adjective modifier. E.g. 合法/a 收入/n (lawful incoming)
  - **PZN** - Noun and its nominal modifier. E.g. 道德/n 标准/n (moral standard)
  - **SBI** - Predicate and its object. E.g. 保护/v 文物/n (protect culture relic)
  - **SBU** - Predicate and its complement. E.g. 医治/v 无效/v (ineffectively treat)
  - **ZZ** - Predicate and its adverbial modifier. E.g. 沉重/ad 打击/v (heavily strike)
  - **SD** - Serial verb constructions. E.g. 跟踪/v 报导/v (trace and report)
  - **ZW** - Predicate and its subject. E.g. 财产/n 转移/v (property transfer)
  - **AA** – Adjective and its adverbial modifier. E.g. 极其/d 惨痛/a (greatly painful)
Pass 1. Concordance and dependent word identification

- Headword concordance

  确保/v 人民/n 群众/n 的/u 生命/n 财产/n 安全/an
  确保/v 长江/ns 安全/an 度汛/v

- Manual identification of syntactically and semantically dependent words surrounding the headword

  <p>确保/v 人民/n 群众/n 的/u 生命/n 财产/n 安全/an</p>
  <depend search=“安全/an ”head=”确保/v” depend =”安全/an” relation =”SBI” ></depend>
  <depend search=“安全/an ”head=”安全/an” depend =”生命/n” relation =”PZN” ></depend>
  <depend search=“安全/an ”head=”安全/an” depend =”财产/n” relation =”PZN” ></depend>

  <p>确保/v 长江/ns 安全/an 度汛/v</p>
  <depend search=“安全/an ”head=”度汛/v” depend =”安全/an” relation =”ZZ” ></depend>
Pass 2. N-gram collocation annotation

- Identify word combinations frequently co-occur in consecutive positions to be extracted as n-gram collocations.
- No further analysis on the internal syntactic and semantic information is carried out.

<p>确保/v 人民/n 群众/n 的/u 生命/n 财产/n 安全/an</p>

<ncolloc search="""" an"""">安全/an 生命/n 财产/n 安全/an""""</ncolloc>
Pass 3. Bi-gram collocations annotation

- All two-word combinations are considered candidates
- Bi-gram collocations are type labeled according to manual judgment and the following statistical features
  
  **Strength:** Reflects the co-occurrence frequency significance  
  **Spread:** Reflects the co-occurrence distribution significance  
  **Synonym Substitution Rate:** Measures the substitutability  
  **Distribution Similarity:** Measures the distribution similarity between a collocation candidate and the statistically expected distribution.

What is the following related to the above statements?

```xml
<bcolloc search="安全/an" col="确保/v" head=“确保/v” type=”2” relation=”SBI”></bcolloc>
<bcolloc search="安全/an” col=”度汛/v” head=“度汛/v” type=”3” relation=”ZZ”></bcolloc>
```
Achievements

- From the collocation bank, 23,581 identical bi-gram collocations are identified which is called *PolyU Collocation Collection (PCC)*.
  - “The Dictionary of Modern Chinese Collocation” provides 35,742 typical collocations for these headwords, which is called *Mei’s Collocation Collection (MCC)*
  - There are 19,967 common entries in PCC and MCC, which indicates good linguistic consistency
  - 3,614 new collocations are obtained that are not recorded in MCC, which means collocation bank is helpful to enrich collocation dictionary

- Collocation bank provide accurate statistics
  - The manual identification bi-gram collocations and verification of co-occurrence are helpful
  - Statistical information collected from collocation bank is more accurate and can be more useful to linguistic research, and it is essential to improving the automatic collocation extraction systems
A Unicode-based Adaptive Segmenter

Qin Lu\textsuperscript{1}, Shiu-tong Chan\textsuperscript{1}, Baoli Li\textsuperscript{2} and Shiwen Yu\textsuperscript{2}

\textsuperscript{1}Department of Computing, The Hong Kong Polytechnic University
\textsuperscript{2}Institute of Computational Linguistics, Beijing University
Design Principles

- No restriction to one text coding: Handle both traditional/simplified characters
  
  **Example:** I need to go to school the day after tomorrow

  - Simplified: [我][后日][要][上学]
  - Traditional: [我][後日][要][上學]
  - Mixed: [我][後日][要][上学]

- Output in chosen form

- Reusable and adaptive: Modular design: divide and conquer
  
  - Training of terms in traditional/Taiwan/HK as additional pluggable dictionaries
System Architecture

Segmenter
- Pre-Processor
- Kernel
- Post Processor

Data manager
- Word Extractor
- Dictionary management
Basic System Functions

- Pre-processing: Dealing with Eng-Chinese separation
- Segmentation:
  - Dictionary based with word frequency and conditional probability
  - Chinese name recognizer
  - PoS tagging (optional) based on Peking Univ. standard
## Word Extractor

### HK Unique terms

<table>
<thead>
<tr>
<th>Bigram</th>
<th>count</th>
<th>freqforward</th>
<th>freqbackward</th>
</tr>
</thead>
<tbody>
<tr>
<td>差餉</td>
<td>818</td>
<td>0.061425</td>
<td>0.943483</td>
</tr>
<tr>
<td>入伙</td>
<td>734</td>
<td>0.007373</td>
<td>0.583002</td>
</tr>
<tr>
<td>叱吒</td>
<td>120</td>
<td>0.902256</td>
<td>0.794702</td>
</tr>
<tr>
<td>蜃殼</td>
<td>106</td>
<td>0.751773</td>
<td>0.120045</td>
</tr>
<tr>
<td>叮噹</td>
<td>44</td>
<td>0.330827</td>
<td>0.473118</td>
</tr>
<tr>
<td>普洱</td>
<td>44</td>
<td>0.003138</td>
<td>0.483516</td>
</tr>
<tr>
<td>磅礴</td>
<td>34</td>
<td>0.012523</td>
<td>1.000000</td>
</tr>
<tr>
<td>孵童</td>
<td>15</td>
<td>1.000000</td>
<td>0.001848</td>
</tr>
<tr>
<td>菓菓</td>
<td>15</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>鏗絲</td>
<td>14</td>
<td>0.933333</td>
<td>0.003801</td>
</tr>
<tr>
<td>鯈魚</td>
<td>9</td>
<td>1.000000</td>
<td>0.002222</td>
</tr>
<tr>
<td>忤逆</td>
<td>7</td>
<td>1.000000</td>
<td>0.002152</td>
</tr>
</tbody>
</table>
## Performance Evaluation: All 4 data sets

<table>
<thead>
<tr>
<th>Data</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>$R_{ooV}$</th>
<th>$R_{iv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>0.892</td>
<td>0.853</td>
<td>0.872</td>
<td>0.236</td>
<td>0.906</td>
</tr>
<tr>
<td>CTB</td>
<td>0.853</td>
<td>0.806</td>
<td>0.829</td>
<td>0.578</td>
<td>0.914</td>
</tr>
<tr>
<td>HK</td>
<td>0.909</td>
<td>0.863</td>
<td>0.886</td>
<td>0.579</td>
<td>0.935</td>
</tr>
<tr>
<td>PK</td>
<td>0.94</td>
<td>0.911</td>
<td>0.925</td>
<td>0.647</td>
<td>0.962</td>
</tr>
</tbody>
</table>