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**

|                        | Type 0 Idiomatic Collocation | Type 1 Fixed Collocation | Type 2 Strong Collocation | Type 3 Loose Collocation |
|------------------------|------------------------------|--------------------------|---------------------------|--------------------------|
| Compositional          | No                           | Limited to yes           | Yes                       | Yes                      |
| Synonym substitutable  | No                           | No                       | Very limited              | Limited                  |
| Order alter            | No                           | No                       | Yes                       | Yes                      |
| Modifiable             | No                           | No                       | Very Limited              | Limited                  |
| Statistic significance | Not required                 | Not required             | Required                  | Strongly Required        |
| Examples               | 纲木求鱼 (climb the tree to get fish – wrong method) 蓝牙 (Bluetooth) | 外交/n 豁免权/n(diplomat immunity) | 缔结/v 同盟/n 联盟/n(form alliances) | 合法/v 收入/n 正当/v 收入/n 合法/v 收益/n (lawful income) |
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="安全/an" w1="生命/n" w2="财产/n" w3="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 Lu¹, Shiu-tong Chan¹, Baoli Li² and Shiwen Yu²

¹Department of Computing, The Hong Kong Polytechnic University
²Institute 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

| Bigram  | count | freqforward | freqbackward |
|---------|-------|-------------|--------------|
| 差餉   | 818   | 0.061425    | 0.943483     |
| 入伙   | 734   | 0.007373    | 0.583002     |
| 叱吒   | 120   | 0.902256    | 0.794702     |
| 螵殼   | 106   | 0.751773    | 0.120045     |
| 叮噹   | 44    | 0.330827    | 0.473118     |
| 普洱   | 44    | 0.003138    | 0.483516     |
| 磅礡   | 34    | 0.012523    | 1.000000     |
| 異童   | 15    | 1.000000    | 0.001848     |
| 萱薀   | 15    | 1.000000    | 1.000000     |
| 錆絲   | 14    | 0.933333    | 0.003801     |
| 鰤魚   | 9     | 1.000000    | 0.002222     |
| 忤逆   | 7     | 1.000000    | 0.002152     |
Performance Evaluation: All 4 data sets

| Data | R   | P   | F   | R_{oo} | R_{iv} |
|------|-----|-----|-----|--------|--------|
| AS   | 0.892 | 0.853 | 0.872 | 0.236  | 0.906  |
| CTB  | 0.853 | 0.806 | 0.829 | 0.578  | 0.914  |
| HK   | 0.909 | 0.863 | 0.886 | 0.579  | 0.935  |
| PK   | 0.94  | 0.911 | 0.925 | 0.647  | 0.962  |