Mapping and Generating Classifiers using an Open Chinese Ontology

Luis Morgado Da Costa, Francis Bond and Helena Gao

- Introduction
  - Motivation
    - Previous Work
      - Our Algorithm
    - Future Work
What is a classifier?

(measure/counter word)

(a slice of cake)

(sān zhī lǎo shǔ)

(三只老鼠)

(three mice)

(yí piàn dàn gāo)

(一片蛋糕)
What is a classifier?
(measure/counter word)

Word or morpheme that some languages require (or allow) in the quantification of noun phrases.

And while, semantically, they do not introduce a referent or event, they impose/are restricted by something in the referent.

semantic features
Types of Classifiers

There are many types of classifiers: (Bond and Paik, 2000)

- **sortal** (which classify the kind of the noun phrase they quantify);
- **event** (which are used to quantify events);
- **mensural** (which are used to measure the amount of some property);
- **group** (which refer to a collection of members);
- **taxonomic** (which force the noun phrase to be interpreted as a generic kind)

Most languages make use of some / different types of classifiers

- a kilo of coffee (mensural classifier)
- a school of fish (group classifiers)
- a head of cattle / a loaf of bread (? traces of sortal classifiers)
Sortal Classifiers

- A wheel, a block, a wedge or a brick of cheese?
Sortal Classifiers

- **A wheel, a block, a wedge or a brick of cheese?**
  - It depends on the shape of the cheese!
Examples (Mandarin Chinese)

(1) 两 只 狗
liǎng zhī gǒu
2  CL  dog
“two dogs”

(2) 两条狗
liǎng tiáo gǒu
2  CL  dog
“two dogs”

(3) 两条路
liǎng tiáo lù
2  CL  road
“two roads”

(4) 三 台 电脑
sān tái diànnǎo
3  CL  computer
“three computers”

(5) *三 只 电脑
sān zhī diànnǎo
3  CL  computer
“three computers”
Many NLP tasks need these resources:

- Machine Translation

```
sān zhī lǎo shǔ
三只老鼠
(three mice)
```

- Language Learning (CLs are hard for L2 learners of Mandarin)

- Word Sense Disambiguation
The overlap of semantic features can help WSD tasks

- 一个 木头
  yī ge mùtou
  1 CL log (of wood) / blockhead
  “a log / blockhead”

- 一位 木头
  yī wèi mùtou
  1 CL blockhead
  “a blockhead”

- 一根 木头
  yī gēn mùtou
  1 CL log (of wood)
  “a log”

(general classifier)
(human, formal classifier)
(long, slender objects classifier)
In Chinese, Sortal Classifier (S-CL) usage is complex and mandatory!
(many-to-many relations between nouns and classifiers, with different levels of acceptability depending on shape, size, function, etc.)

No machine tractable, open resources describing S-CL usage…
(Many paper resources exist, but they focus more on what kind of nouns can be used with a particular classifier)

Producing an exhaustive list of noun-classifiers is impossible!
(Nouns are open class words)
Mapping and Generating Classifiers using an Open Chinese Ontology

Automatically
Previous Work...

- The first theoretical description of leveraging hierarchical semantic classes to generalize noun-CL pairs; (Sornlertlamvanich et al., 1994) (for Thai, produced no living results)

- Bond and Paik (2000) and Paik and Bond (2001) further develop these ideas to develop similar works for Japanese and Korean. (similar works for Japanese and Korean, hand rules to propagate through Goi-Taikei (and CorNet); achieve up to 81% of generation accuracy)

- Mok et al. (2012) develop a similar approach using the Japanese Wordnet and the Chinese Bilingual Wordnet (Report a generation score of 78.8% and 89.8%, over a small news corpus)
We wanted to mimic this mapped noun-CL pairs:

- Fully automated extractive and mapping algorithm
- Mapping to Chinese Open Wordnet (COW) (Wang and Bond, 2013)

Non-ranked
- top-down propagation (noisy)
- Low coverage (too much human work)
Enriching COW with S-CLs

The integration between corpora and knowledge rich resources, like dictionaries, can offer good insights and generalisations on linguistic knowledge. (Huang et al., 1998)

- **Chinese Open Wordnet (COW)** (Wang and Bond, 2013)
  
  Large open, machine tractable, Chinese semantic ontology
  
  + Bilingual Ontological Wordnet (BOW) + Southeast University Wordnet (SEW) + Wiktionary and CLDR data (Extended OMW)
  
  (261k nominal lemmas, from which over 184k were unambiguous)

- **Chinese Corpora** (Sentence delimited, segmented, POS tagged)
  
  
  (approx. 30 million sentences, 950 million words)
  
  Google Ngram corpus for Chinese, 2012

- **A list of 204 Chinese S-CLs** (Huang et al., 1997)
But… extracting noun-CL pairs from corpora is not straightforward:

- **Long distance dependencies**

  *The book that was bought by those three students in that old bookstore.*
  
  [ *that CLASSIFIER … … … … book* ]

- **Anaphoric or deictic references**

  *I prefer this.*  
  (omitting the referent)
  
  [ *I prefer this* CLASSIFIER ]

- **Synecdoches [at least in Japanese]**

  *Those 2 pizzas are very friendly.*  
  (referring to the customers who ordered them)
  
  [ *Those 2 HUMAN-CLASSIFIER pizzas are very friendly* ]
Our Work

- Two S-CL dictionaries (w/ frequency information):
  - lemma based dictionary (independent from COW)
  - concept based dictionary (COW)

- Our Algorithm:
  - Extracting Classifier-Noun Pairs
  - Map to COW & Extend coverage
  - Automatic Evaluation (80% Training + 10% Development + 10% Evaluation)
Matching **very restrictive** POS patterns of the form:

(determiner or numeral) + (CL) + (noun) + (end of sentence punctuation/select conjunctions)

This filters out long dependencies after the CL, and tries to **maximally reduce** the noise introduced by anaphoric and deictic uses of CLs. [helpless against synecdoches]

**(CL) + (noun) pairs**

- Feed the lemma based dictionary
- Frequency information is also stored (used in generation)
- Training Set: 435k + 13.5M (Google Ngrams) noun-CL tokens pairs
Lemma Dictionary

类别 (lèibié) “category”

✔ 58: 个 ge
✔ 1: 项 xiàng

养鸡场 (yǎngjīchǎng) “chicken farm”

✔ 6: 个 ge
✔ 3: 家 jiā
✔ 1: 座 zuò

× 2: 只 zhǐ

Some noise...
+ missing 间 jiān and 所 suǒ

SPOILER ALERT!

只 zhǐ can be used with 养鸡 (yǎngjī)
Mapping S-CLs to COW

- Map unambiguous lemmas to COW
  (i.e. that belong to a single concept)

- Frequency information and possible CLs are stacked for each matched sense. (i.e. store the union of all senses)

类别 (lèibié) “category”  >>>  ID 05838765-n “a general concept that marks divisions or coordinations in a conceptual scheme”

✔️ 58: 个 ge
✔️ 1: 项 xiàng

+ data from 范畴 (fànchóu)
+ data from 种类 (zhǒnglèi)

✔️ 132: 个 ge
✔️ 2: 项 xiàng
Extend COW’s Coverage

- **Principle:**
  
  Wordnets should be able, to some extent, to **model the semantic features hierarchy** that link nouns and CLs.

  For every concept with CL data:
  - Search 10 levels of hypernymy and hyponymy
  - If a CL match is found, **share it**!
  - Sums frequencies of all matches

  **We do not blindly assign CLs down the concept hierarchy**, making it depend on previously extracted information for both hypernyms and hyponyms.
Automatic Evaluation

(Dev-Set = 37.4k & Test-Set = 39.9k tokens of noun-CL pairs)

- We evaluated on an automated task of CL prediction & generation
  (i.e. trying to predict if a classifier is valid + matching with the most freq. CL)

- Dev-set (10% of the data) was used to filter data by frequency
- T frequency: from 1 to 5 minimum frequency to be considered
- Best T was tested, again, against the test-set (10% of the data)

- Baseline: assigning $\uparrow$ (ge) as the only CL for every entry
- Fallback: always assigning $\uparrow$ (ge) as a possible CL
Results

- Concept mapping wins the prediction of the validity of a CL (wn-unamb);

- Lemma mapping wins in the generation task (lemma-unamb-mfcl); this was unexpected!

- Filtering didn’t help performance… Not enough data! But...

- The coverage of the concept dictionary reduces much less drastically (x2.25 senses per concept)

- Also, the increase in no-info is larger than the decrease in performance

- Filtering reduces over-generation (validated but not presented)
Results - Explained

- Why is the concept mapping is outperformed in generation?
  - Incorrect / incomplete concept hierarchy (?)
  - CLs relate better to specific senses than to concepts (?)
  - Noise in the testing data (?) [We don’t yet have a gold set]

- So we went, checked a small sample, and...
  - Found a lots false positives on the lemma mapping introduced also by the lemmatisation and POS tagging errors.

    Roughly **7.5% of invalid lemmas** (i.e. non-words, non-nouns)

  - Mapping to COW filters all (most) invalid lemmas! (they fail to map!)
  - Human checking verified that the concept mapping outperformed the lemma based mapping: **87% vs 76%**
Future Work

- More error analysis
- Create a Gold test-set
- Repeat with more data!
  (e.g. a very large web-crawled corpus)
- Repeat similar approach with other languages (i.e. Japanese)
  (for the most part this approach is language independent)
- Be less naive…
  (Include a measure of Mutual Information, play with vector spaces, etc.)
- Use WSD (e.g. UKB, cross-lingual WSD)
  (and include S-CL mapping of ambiguous senses)
Future Work

- **CLs in Wordnet**
  - ‘x’ as part-of-speech
  - definition with the form “a ... classifier used ..., such as ...”
  - domain usage: classifier (06308436-n)

- **87 Chinese S-CLs in COW**

- **30 Indonesian S-CLs in WN Bahasa**

```
[80000003-x
 lemmas  把 (bā)
definition a sortal classifier used with tools and objects with a handle, such as a hammer, a broom, a guitar or a teapot
domain usage 06308436-n (classifier)
```

```
[80000004-x
 lemmas  根 (gēn)
definition a sortal classifier used for long slender objects, such as a banana, a pillar, a sausage or a needle
domain usage 06308436-n (classifier)
```
## Data: mappings

<table>
<thead>
<tr>
<th>Code</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>07772274-n</td>
<td>颗</td>
<td>1</td>
</tr>
<tr>
<td>14377617-n</td>
<td>个</td>
<td>4</td>
</tr>
<tr>
<td>00429322-n</td>
<td>片</td>
<td>3</td>
</tr>
<tr>
<td>00429322-n</td>
<td>分</td>
<td>2</td>
</tr>
<tr>
<td>00429322-n</td>
<td>份</td>
<td>5</td>
</tr>
<tr>
<td>00429322-n</td>
<td>起</td>
<td>1</td>
</tr>
<tr>
<td>00429322-n</td>
<td>家</td>
<td>1</td>
</tr>
<tr>
<td>00429322-n</td>
<td>丝</td>
<td>1</td>
</tr>
<tr>
<td>07231294-n</td>
<td>次</td>
<td>32</td>
</tr>
<tr>
<td>07231294-n</td>
<td>番</td>
<td>3</td>
</tr>
<tr>
<td>07231294-n</td>
<td>重</td>
<td>4</td>
</tr>
<tr>
<td>07231294-n</td>
<td>个</td>
<td>109</td>
</tr>
<tr>
<td>邮展</td>
<td>次</td>
<td>7</td>
</tr>
<tr>
<td>邮展</td>
<td>届</td>
<td>1</td>
</tr>
<tr>
<td>醒</td>
<td>个</td>
<td>480</td>
</tr>
<tr>
<td>减费</td>
<td>项</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>套</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>篇</td>
<td>11</td>
</tr>
<tr>
<td>小说</td>
<td>集</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>部</td>
<td>69</td>
</tr>
<tr>
<td>小说</td>
<td>个</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>名</td>
<td>2</td>
</tr>
<tr>
<td>小说</td>
<td>本</td>
<td>42</td>
</tr>
<tr>
<td>小说</td>
<td>卷</td>
<td>4</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
<td>Frequency</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>07772274-n</td>
<td>颗</td>
<td>1</td>
</tr>
<tr>
<td>14377617-n</td>
<td>个</td>
<td>4</td>
</tr>
<tr>
<td>00429322-n</td>
<td>片</td>
<td>3</td>
</tr>
<tr>
<td>00429322-n</td>
<td>分</td>
<td></td>
</tr>
<tr>
<td>00429322-n</td>
<td>小说</td>
<td>1</td>
</tr>
<tr>
<td>00429322-n</td>
<td>家</td>
<td>1</td>
</tr>
<tr>
<td>00429322-n</td>
<td>小说</td>
<td>1</td>
</tr>
<tr>
<td>07231294-n</td>
<td>次</td>
<td>32</td>
</tr>
<tr>
<td>07231294-n</td>
<td>番</td>
<td>3</td>
</tr>
<tr>
<td>07231294-n</td>
<td>重</td>
<td>4</td>
</tr>
<tr>
<td>07231294-n</td>
<td>个</td>
<td>109</td>
</tr>
<tr>
<td>邮展</td>
<td>次</td>
<td>7</td>
</tr>
<tr>
<td>邮展</td>
<td>届</td>
<td>1</td>
</tr>
<tr>
<td>醒</td>
<td>个</td>
<td>480</td>
</tr>
<tr>
<td>减费</td>
<td>项</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>套</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>册</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>集</td>
<td>1</td>
</tr>
<tr>
<td>小说</td>
<td>部</td>
<td>69</td>
</tr>
<tr>
<td>小说</td>
<td>本</td>
<td>42</td>
</tr>
<tr>
<td>小说</td>
<td>卷</td>
<td>4</td>
</tr>
</tbody>
</table>
Thank You!