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

NANYANG TECHNOLOGICAL UNIVERSITY

What is a classifier?

(measure/counter word)

(three mice)

(a slice of cake)

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

NANYANG TECHNOLOGICAL UNIVERSITY

Types of Classifiers

- **<u>There are many types of classifiers:</u>** (Bond and Paik, 2000)

  <u>sortal</u> (which classify the kind of the noun phrase they quantify);

  <u>event</u> (which are used to quantify events);

  <u>mensural</u> (which are used to measure the amount of some property);

  <u>group</u> (which refer to a collection of members);

  <u>taxonomic</u> (which force the noun phrase to be interpreted as a generic kind)

- **<u>Most languages make use of some / different types of classifiers</u>**

  - 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

- **<u>A wheel, a block, a wedge or a brick of cheese?</u>**

  It depends on the shape of the cheese!

Examples (Mandarin Chinese)

(1)	两	只	狗
	liǎng	zhǐ	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	zhǐ	diànnǎo
	3	CL	computer

"three computers"

Motivation

- **<u>Many NLP tasks need these resources:</u>**
  - Machine Translation

(three mice)

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

NANYANG TECHNOLOGICAL UNIVERSITY

Classifiers & WSD

- **The overlap of semantic features can help WSD tasks**

  - 一 个 木头 (general classifier)
    yī ge mùtou
    1 CL log (of wood) / blockhead

    "a log / blockhead"

  - 一 位 木头 (human, formal classifier)
    yī wèi mùtou
    1 CL blockhead

    "a blockhead"

  - 一 根 木头 (long, slender objects classifier)
    yī gēn mùtou
    1 CL log (of wood)

    "a log"

Motivation (II)

- **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)

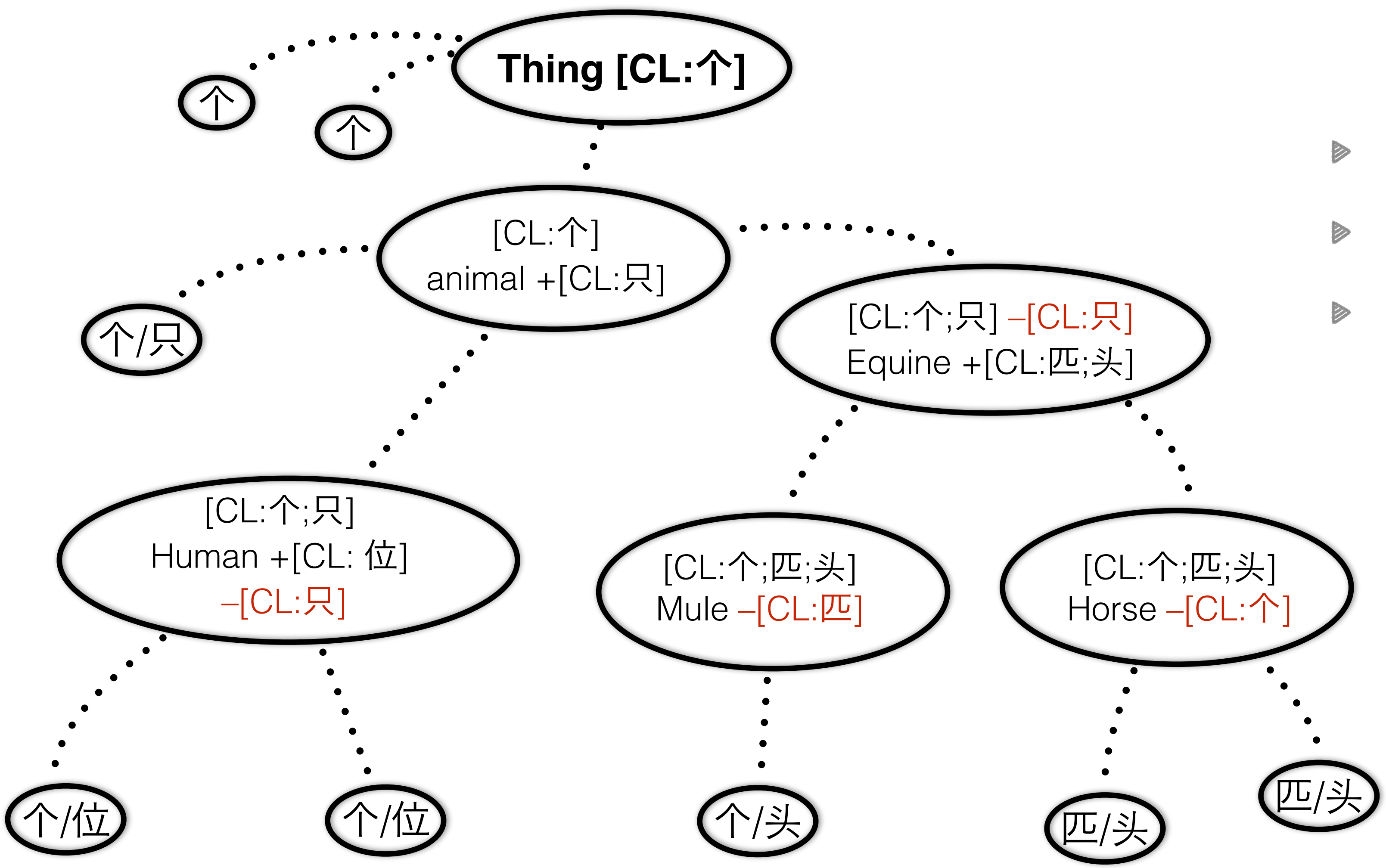
- **Bond and Paik (2000) and Paik and Bond (2001) further develop these ideas to develop similar works for Japanese and Korean.**

  (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)

Mok et al. (2012)



- Non-ranked
- top-down propagation (noisy)
- Low coverage

  (too much human work)

- **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)

NANYANG TECHNOLOGICAL UNIVERSITY

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)

  Chinese Wikipedia, 2nd Edition Chinese Gigaword Corpus, UM-Corpus

  (approx. 30 million sentences, 950 million words)

  Google Ngram corpus for Chinese, 2012

- **A list of 204 Chinese S-CLs** (Huang et al., 1997)

Problems in Automated Approches

- **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)

Extracting Classifier-Noun Pairs

- **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

NANYANG TECHNOLOGICAL UNIVERSITY

Lemma Dictionary

类别 *(lèibié)* "category"

- ✔ 58: 个 *ge*
- ✔ 1: 项 *xiàng*

养鸡场 *(yǎngjīchǎng)* "chicken farm"

- ✔ 6: 个 *ge*
- ✔ 3: 家 *jiā*
- ✘ 2: 只 *zhǐ*
- ✔ 1: 座 *zuò*

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)*

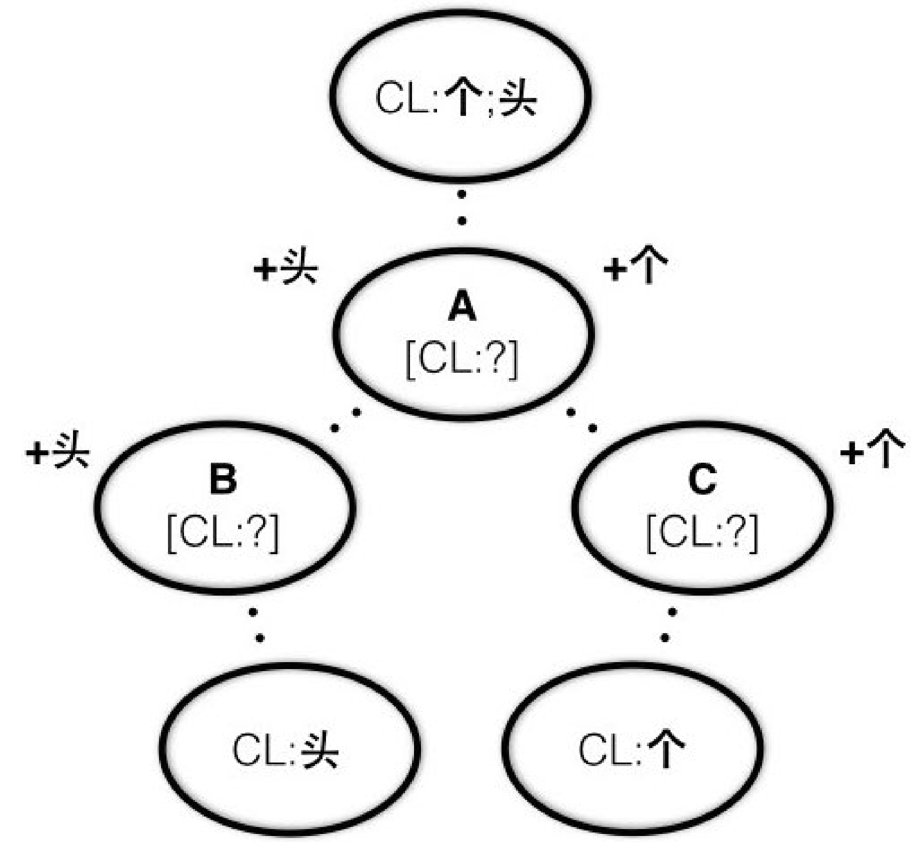
✔ 132: 个 *ge*

✔ 2: 项 *xiàng*

Extend COW's Coverage

- **Principle:**

  Wordnets should be able, to some extent, to <u>model the semantic features hierarchy that link nouns and CLs</u>.



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.

NANYANG TECHNOLOGICAL UNIVERSITY

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 个 *(ge)* as the only CL for every entry
- **Fallback:** always assigning 个 *(ge)* as a possible CL

Results

	$\tau = 1$	$\tau = 3$	$\tau = 5$	*Test*
baseline	44.2	44.2	44.2	40.4
All lemmas				
lem-all	92.7	88.5	86.2	93.6
lem-all-mfcl	75.1	73.8	72.8	78.9
lem-all-no-info	4.7	9.2	12.1	4.1
Unamb. lemmas				
lem-unamb	93.2	88.2	85.5	94.5
wn-unamb	**95.1**	90.9	88.3	**95.9**
lem-unamb-mfcl	**77.0**	75.5	74.1	**77.9**
wn-unamb-mfcl	72.3	71.6	70.7	73.5
lem-unamb-no-info	3.4	9.5	13.6	2.8
wn-unamb-no-info	1.7	5.3	8.3	1.5
Coverage				
lemmas-w/cl	32.4k	10.4k	7.0k	
wn-concepts-w/cl	22.7k	15.0k	12.3k	

- 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 (bǎ)
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

07772274-n	颗	1
14377617-n	个	4
00429322-n	片	3
00429322-n	分	2
00429322-n	份	5
00429322-n	起	1
00429322-n	家	1
00429322-n	丝	1
07231294-n	次	32
07231294-n	番	3
07231294-n	重	4
07231294-n	个	109

邮展	次	7
邮展	届	1
醒	个	480
减费	项	1
小说	套	1
小说	篇	11
小说	集	1
小说	部	69
小说	个	1
小说	名	2
小说	本	42
小说	卷	4

Data: mappings

07772274-n	颗	1
14377617-n	个	4
00429322-n	片	3
00429322-n	方	[illegible]
00429322-n	[illegible]	[illegible]
00429322-n	[illegible]	1
00429322-n	家	1
00429322-n	[illegible]	[illegible]
07231294-n	次	32
07231294-n	番	3
07231294-n	重	4
07231294-n	个	109

邮展	次	7
邮展	届	1
醒	个	480
减费	项	1
小说	套	1
小说	[illegible]	1
小说	集	[illegible]
小说	部	[illegible]
小说	[illegible]	1
小说	名	2
小说	本	42
小说	卷	4

Thank You!

NANYANG
TECHNOLOGICAL
UNIVERSITY