Analogical Reasoning on Chinese Morphological and Semantic Relations

Shen Li ♠ Zhe Zhao Renfen Hu ♠ Wensi Li Tao Liu Xiaoyong Du
♠ {shen, irishu}@mail.bnu.edu.cn Beijing Normal University & Renmin University of China

GitHub: Chinese-Word-Vectors
Over 2,000

Scan me

Introduction

Given the word representations, analogy questions can be automatically solved via vector computation:

$$apples - apple + car \approx cars \quad \text{(morphological)}$$
$$king - man + woman \approx queen \quad \text{(semantic)}$$

It is well known that linguistic regularities vary a lot among different languages. For example, Chinese is a typical **analytic language** which lacks inflection.

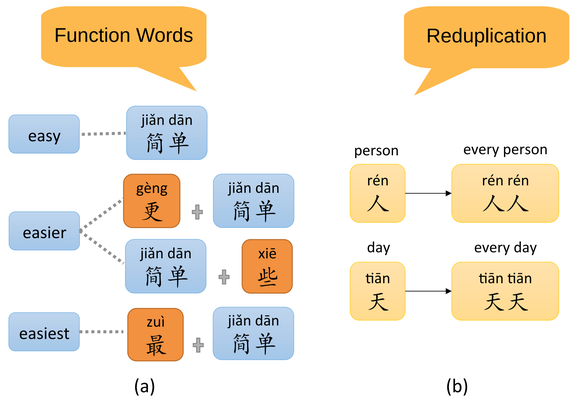

Figure 1: Examples of Chinese lexical knowledge.

Semantic Relations

We present **28 semantic relations** in 4 aspects.

Geography	country-capital, country-currency, province-abbreviation, province-capital, province-dramma, etc.
History	dynasty-emperor, dynasty-capital, title-emperor, celebrity-country
Nature	number, time, animal, plant, body, physics, weather, reverse, color, etc.
People	finding-scientist, work-writer, family members, etc.

Morphological Relations

• Reduplication

Reduplication means a morpheme is repeated to form a new word, which is semantically and/or syntactically distinct from the original morpheme.

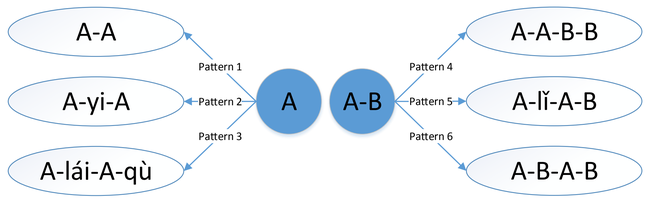

Figure 2: Reduplication patterns of A and A-B.
(A and B are distinct morphemes.)

Taking A→AA as an example:

- bà (dad) → bà-bà (dad)
- tiān (day) → tiān-tiān (everyday)
- shuō (say) → shuō-shuo (say a little)
- kàn (look) → kàn-kàn (have a brief look)
- dà (big) → dà-dà (very big; greatly)
- shēn (deep) → shēn-shēn (deeply)

• Semi-affixation

Since Chinese is a typical isolating language that has few affixes, we describe the similar morphological process with semi-affixation suggested by Liu et al. 2001. To model the semi-affixation process, we uncover **21 semi-prefixes** and **41 semi-suffixes**.

Taking “dì-” and “-zi” as examples:

- yī (one) → dì-yī (first)
- èr (two) → dì-èr (second)
- pàng (fat) → pàng-zi (a fat man)
- shòu (thin) → shòu-zi (a thin man)

Pre-trained Embeddings

This project provides **100+ Chinese Word Vectors (embeddings)** trained with different representations (**dense and sparse**), context features (**word, ngram, character, and more**), and corpora.

Corpus	Size
Baidu Encyclopedia 百度百科	4.1G
Wikipedia_zh 中文维基百科	1.3G
People's Daily News 人民日报	3.9G
Sogou News 搜狗新闻	3.7G
Financial News 金融新闻	6.2G
Zhihu_QA 知乎问答	2.1G
Weibo 微博	0.73G
Literature 文学作品	0.93G
Mixed-large 综合	22.6G
Complete Library in Four Sections 四库全书	1.5G

Table 1: Corpus.

Feature	Co-occurrence Type
Word	Word → Word
Ngram	Word → Ngram (1-2)
	Word → Ngram (1-3)
	Ngram (1-2) → Ngram (1-2)
Character	Word → Character (1)
	Word → Character (1-2)
	Word → Character (1-4)
Radical	Radical
Position	Word → Word (left/right)
	Word → Word (distance)
Global	Word → Text
Syntactic Feature	Word → POS
	Word → Dependency

Table 2: Various Co-occurrence Information.

*All text data are preprocessed by removing HTML and XML tags. Only the plain text are kept and HanLP(v_1.5.3) is used for word segmentation.

CA8 Dataset

CA8 contains **17813** analogy questions and covers comprehensive morphological and semantic relations. CA8-morphological (**CA8-Mor**) contains **10177** morphological questions based on two types of relations: reduplication and semi-affixation. CA8-semantic (**CA8-Sem**) contains **7636** semantic questions divided into 4 categories and 28 sub-categories.