Chinese Analogical Reasoning on Morphological and Semantic Relations

Anonymous ACL submission

A Appendix

CA8 incorporates comprehensive morphological and semantic relations in Chinese. Specifically, CA8-morphological (CA8-Mor) contains 10177 morphological questions, which are constructed based on two types of relations: reduplication and semi-affixation. Table 1 and Table 2 respectively list the details and examples of these two types of relations. CA8-semantic (CA8-Sem) contains 7636 semantic questions, which can be divided into 4 categories and 28 sub-categories. Table 3 shows the detailed description of each semantic relation.

We make great efforts to collect and build corpora of different sizes and domains. Table 4 lists the information of the six corpora used in our experiments. All the text data are preprocessed via the following steps:

- Remove the html and xml tags from the texts and set the encoding as utf-8. Digit and punctuation are remained.
- Convert traditional Chinese characters into simplified characters with Open Chinese Convert (OpenCC)[1].
- Conduct Chinese word segmentation with HanLP(v_1.5.3)[2].

Category	Sub-category	POS	Morphological Function	Example
A	A A	Noun	Form kinship terms	爸(dad) → 爸爸(dad)
			Yield every/each meaning	天(day) → 天天(everyday)
		Measure	Yield every/each meaning	个(-) → 个个(every/each)
		Verb	Signal doing something a little bit	说(say) → 说说(say a little)
			Signal things happen briefly	看(look) → 看看(have a brief look)
		Adjective	Intensify the adjective	大(big) → 大大(very big)
			Transform it to adverbs	慢(slow) → 慢慢(slowly)
	A yi A	Verb	Signal trying to do something	吃(eat) → 吃一吃(try to eat)
	A lai A qu	Verb	Signal doing something repeatedly	飞(fly) → 飞来飞去(fly around)
AB	A A B B	Noun	Yield many/much meaning	山水(mountain and river) → 山山水水(many mountains and rivers)
		Verb	Indicate a continuous action	说笑(laugh and chat) → 说说笑笑(laugh and chat for a while)
		Adjective	Intensify the adjective	清楚(clear) → 清清楚楚(very clear)
			Yield the meaning of not uniform	大小(size) → 大大小小(all sizes)
		Adverb	Intensify the adverb	彻底(completely) → 彻彻底底(totally and completely)
	A li A B	Adjective	Oralize the adjective and yield derogatory meaning	慌张(flurried) → 慌里慌张(anxious)
	A B A B	Verb	Signal doing something a little bit	注意(pay attention) → 注意注意(pay a little attention)
		Adjective	Intensify the adjective	雪白(white) → 雪白雪白(very white)
			Transform it to a verb	高兴(happy) → 高兴高兴(make someone happy)

Table 1: Detailed information of reduplication relations in CA8.

[1] https://github.com/BYVoid/OpenCC
[2] https://github.com/hankcs/HanLP

Category	Affixoid	Example	Affixoid	Example
Semi-prefix	第	一(one) → 第一(first)	次	大陆(continent) → 次大陆(subcontinent)
	初	一(one) → 初一(the first day of a lunar month)	非	常规(conventional) → 非常规(unconventional)
	十	一(one) → 十一(eleven)	每	次(time) → 每次(every time)
	周	一(one) → 周一(Monday)	全	明星(star) → 全明星(all star)
	星期	一(one) → 星期一(Monday)	伪	君子(gentlemen) → 伪君子(hypocrites)
	老	虎(tiger) → 老虎(tiger)	亚	热带(tropical zone) → 亚热带(sub-tropical zone)
	小	草(grass) → 小草(grass)	洋	酒(wine) → 洋酒(foreign wine)
	大	海(sea) → 大海(large sea)	总	比分(score) → 总比分(total score)
	半	导体(conductor) → 半导体(semiconductor)	反	物质(matter) → 反常规(antimatter)
	单	细胞(cell) → 单细胞(unicell)	副	总统(president) → 副总统(vice president)
	超	链接(link) → 超链接(hyperlink)		
Semi-suffix	们	我(I) → 我们(we)	主义	乐观(optimistic) → 乐观主义(optimism)
	里	这(here) → 这里(here)	鬼	吝啬(stingy) → 吝啬鬼(miser)
	些	这(this) → 这些(these)	式	中(Chinese) → 中式(Chinese style)
	样	这(this) → 这样(such)	队	考古(archaeology) → 考古队(archaeological team)
	个	这(this) → 这个(this one)	色	黄(yellow) → 黄色(the yellow color)
	边	这(this) → 这边(here)	学	地质(geology) → 地质学(discipline of geology)
	种	这(this) → 这种(this kind)	论	宿命(fate) → 宿命论(fatalism)
	次	这(this) → 这次(this time)	站	汽车(bus) → 汽车站(bus station)
	儿	这(this) → 这儿(here)	仪	光谱(spectrum) → 光谱仪(spectrograph)
	部	东(east) → 东部(east)	界	学术(academic) → 学术界(academia)
	中	心(heart) → 心中(in the heart)	族	追星(chasing a star) → 追星族(fans)
	上	山(mountain) → 山上(on the mountain)	棍	赌(gamble) → 赌棍(gambler)
	面	前(front) → 前面(in the front)	灾	雨(rain) → 雨灾(rain disaster)
	者	强(strong) → 强者(the strong one)	气	冷(cold) → 冷气(cold air)
	家	科学(science) → 科学家(scientist)	性	酸(acid) → 酸性(acidic)
	子	胖(fat) → 胖子(a fat man)	厅	歌(song) → 歌厅(karaoke)
	头	木(wood) → 木头(wood)	机	复印(copy) → 复印机(copier)
	工	木(wood) → 木工(carpentry)	法	说(say) → 说法(saying)
	匠	木(wood) → 木匠(carpenter)	剧	粤(Yue) → 粤剧(Cantonese Opera)
	星	笑(laugh) → 笑星(comedian)	长	船(ship) → 船长(captain of a ship)
	手	老(old) → 老手(old hand)		

Table 2: Detailed information of semi-affixation relations in CA8.

Categories	Sub-categories	Examples
Geography	country-capital	中国(China)-北京(Beijing)
	country-currency	中国(China)-人民币(Chinese yuan)
	province-abbreviation	广东(Guangdong)-粤(Yue)
	province-capital	广东(Guangdong)-广州(Guangzhou)
	province-drama	广东(Guangdong)-粤剧(Cantonese Opera)
	province-channel	广东(Guangdong)-广东卫视(Guangdong Satellite TV)
	province-university	浙江(Zhejiang)-浙江大学(Zhejiang University)
	city-university	南京(Nanjing)-南京大学(Nanjing University)
	university-abbreviation	师范大学(Normal University)-师大(Normal University)
History	dynasty-emperor	汉(Han)-刘邦(Liu Bang)
	dynasty-capital	秦(Qin)-咸阳(Xian Yang)
	title-emperor	汉高祖(Emperor Gaozu of Han)-刘邦(Liu Bang)
	celebrity-country	屈原(Qu Yuan)-楚国(Country Chu)
Nature	number	第一(first)-状元(the first in an imperial examination)
	time	春节(Spring Festival)-正月(the first month in a lunar year)
	animal	公鸡(cock)-母鸡(hen)
	plant	杏树(apricot tree)-杏(apricot)
	ornament	手指(finger)-戒指(ring)
	chemistry	盐(salt)-氯化钠(sodium chloride)
	physics	冰(ice)-水蒸气(steam)
	weather	小满(Grain Full)-夏天(summer)
	reverse	松(loose)-紧(tight)
	color	海(sea)-蓝色(blue)
People	company-founder	阿里巴巴(Alibaba)-马云(Ma Yun)
	work-scientist	地动仪(seismograph)-张衡(Zhang Heng)
	work-writer	朝花夕拾(Dawn Blossoms Plucked at Dusk)-鲁迅(Lu Xun)
	family-member	爷爷(grandfather)-孙子(grandson)
	student-degree	小学(elementary school)-小学生(schoolchild)

Table 3: Detailed information of semantic relations in CA8.

Corpus	Size	#tokens	$\|V\|$	Description
Baidubaike	4.3G	745M	271K	Chinese wikipedia data from https://baike.baidu.com/
Wikipedia	1.2G	223M	133K	Wikipedia data obtained from https://dumps.wikimedia.org/
People's Daily News	4.2G	669M	171K	News data from People's Daily(1946-2017) http://data.people.com.cn/
Sogou news	4.0G	657M	176K	News data provided by Sogou Labs http://www.sogou.com/labs/
Zhihu QA	2.2G	384M	123K	Chinese QA data from https://www.zhihu.com/, including 32137 questions and 3239114 answers
Combination	15.9G	2678M	456K	We build this corpus by combining the above corpora

Table 4: Detailed information of the corpora. #tokens denotes the number of tokens in corpus. $|V|$ denotes the vocabulary size.