ANALOGICAL REASONING ON
CHINESE MORPHOLOGICAL AND SEMANTIC RELATIONS

汉语的词向量与类推测试

北京大学 中文信息处理 唐乾桐

ANALOGICAL REASONING ON
CHINESE MORPHOLOGICAL AND SEMANTIC RELATIONS

汉语的词向量与类推测试

北京大学 中文信息处理 唐乾桐

1.问题描述
2.词嵌入概览
3.类推法概览
4.本文的类推实验设计
5.实验结果

问题描述

What & How

问题描述

- 在自然语言处理中，语言单位（字、词、句子等）需要有高效简洁的数据表示（data representation）：
  - 直接承载语言信息，为更高级的语言处理任务创造条件。
  - 节省计算资源
- 词嵌入（词向量）就是词的一种数据表示。
- 词嵌入的追求（理论上）：
  - 直接承载词所蕴含的所有语言信息，包括语义信息和语法信息。
  - 节省越多的计算资源

延伸阅读：数据编码技术

问题描述

```
In [7]: pre_embedding = word2vec.load('data/zi2vec2.bin')
        pre_embedding['爱']

Out[7]: array([ 0.02318641, -0.04527554, -0.00997966,  0.13863735,  0.06228667,
                0.05969576,  0.01610527,  0.1015555 ,  0.00208612, -0.00917488,
               -0.12436734,  0.04270805,  0.07257671, -0.26799196, -0.15483236,
               -0.01004719, -0.18201824, -0.0937647 , -0.04718755, -0.08767666,
                0.04883178,  0.14300512, -0.04847946,  0.01450507, -0.08758005,
               -0.04480667,  0.24508943, -0.07102628,  0.00885286,  0.06377392,
                0.01634579, -0.02787561, -0.0118992 ,  0.07996762, -0.04713449,
                0.0683604 , -0.09318909, -0.00282295,  0.01361588, -0.00660162,
               -0.06802392,  0.14408961, -0.2410495 , -0.05232474, -0.06069665,
                0.05748704, -0.0572319 ,  0.1105217 , -0.08457625, -0.11090387,
               -0.04061745,  0.02523119, -0.01537016, -0.09068317,  0.04371744,
               -0.00903666, -0.02944135,  0.09460841, -0.04217609,  0.02083229,
                0.12145197, -0.19373454,  0.10814972, -0.24881546,  0.00418631,
                0.11732625, -0.0183216 , -0.08110762,  0.06749552, -0.07890873,
               -0.1587238 , -0.0488408 , -0.03276642,  0.1176642 , -0.13074872,
                0.18964118,  0.07707223,  0.13461018, -0.06841851, -0.14427578,
               -0.08032991, -0.01852139,  0.12190777,  0.02320155,  0.06564564,
               -0.15792845, -0.09926453, -0.02505823,  0.06297594,  0.22951928,
                0.06359901, -0.0115441 , -0.10704326,  0.12146004, -0.05930306,
               -0.16557473, -0.09982823, -0.00469986,  0.03039106, -0.12550074])
```

- 如何检测词嵌入蕴含了多少语言信息？如何量化？
  - 类推法（word analogy）

词嵌入概览

Word Embedding

词嵌入概览

1. 词向量的本质：

- 词的一种数据表示（Data representation）
- 好的数据表示应该：
  - 便于计算机处理
  - 直接蕴含数据的内部特性
    - 可以通过计算，提取出语义、语法等信息。

延伸阅读：数据编码技术

词嵌入概览

2. 获得词向量的常见方法：

① Count (Sparse)

① PPMI

② Predict (Dense)

① C&W

② CBOW & Skipgram (word2vec)

③ Transformers (Bert)

Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 238-247.

词嵌入概览

① Count：

	I	$love$	$playing$	$football$	$tennis$	you
I	0	2	0	0	0	0
$love$	2	0	3	0	0	1
$playing$	0	3	0	2	1	0
$football$	0	0	2	0	0	0
$tennis$	0	0	1	0	0	0
you	0	1	0	0	0	0

词嵌入概览

① Count：

- 选取上下文

词-文档矩阵、词-词矩阵、n元词组

- 确定矩阵中元素的值

词频、TF-IDF、PMI

- 矩阵分解

高维稀疏-->低纬稠密 / SVD、NMF、CCA、HPCA

词嵌入概览

① Count：PPMI

- 点间互信息(PMI)

$$PMI(w,c) = \log \frac{\hat{p}(w,c)}{\hat{p}(w) \cdot \hat{p}(c)} = \log \frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)}$$

词嵌入概览

① Count：PPMI

- 点间互信息(PMI)

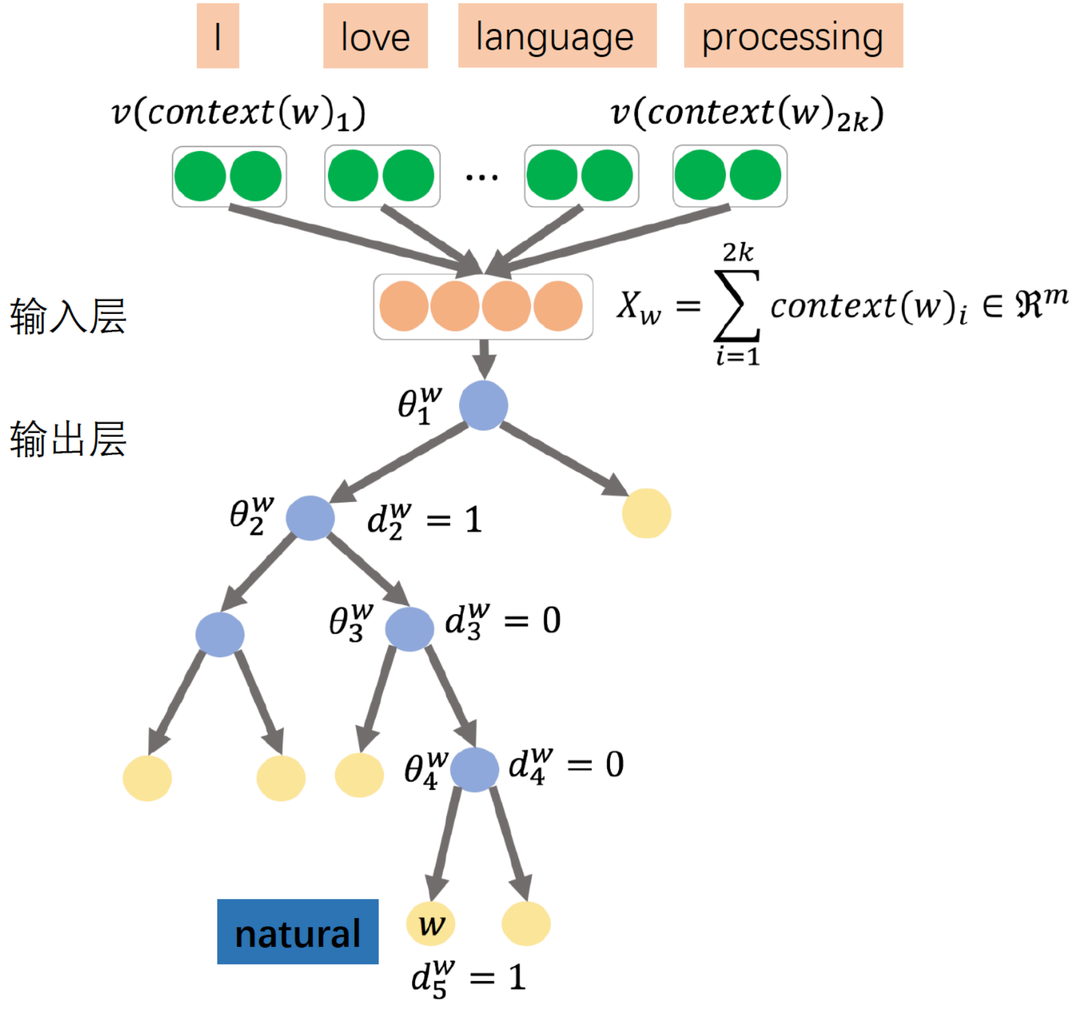
$$PMI(w,c) = \log \frac{\hat{p}(w,c)}{\hat{p}(w) \cdot \hat{p}(c)} = \log \frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)}$$

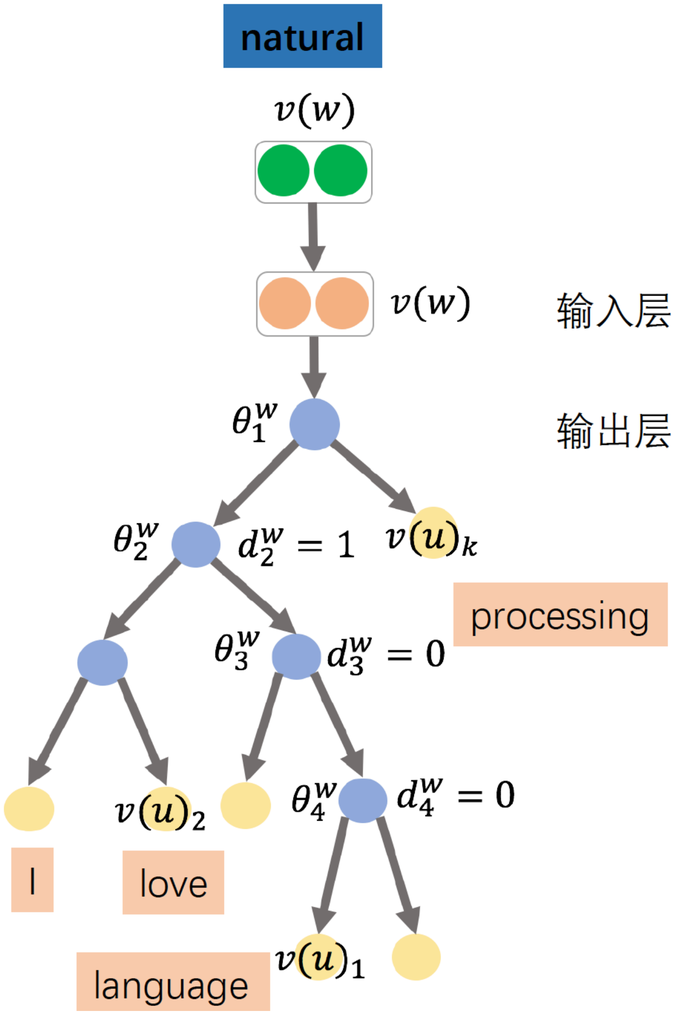
- 正点间互信息(PPMI)

$$PPMI(w,c) = \max(PMI(w,c), 0)$$

词嵌入概览

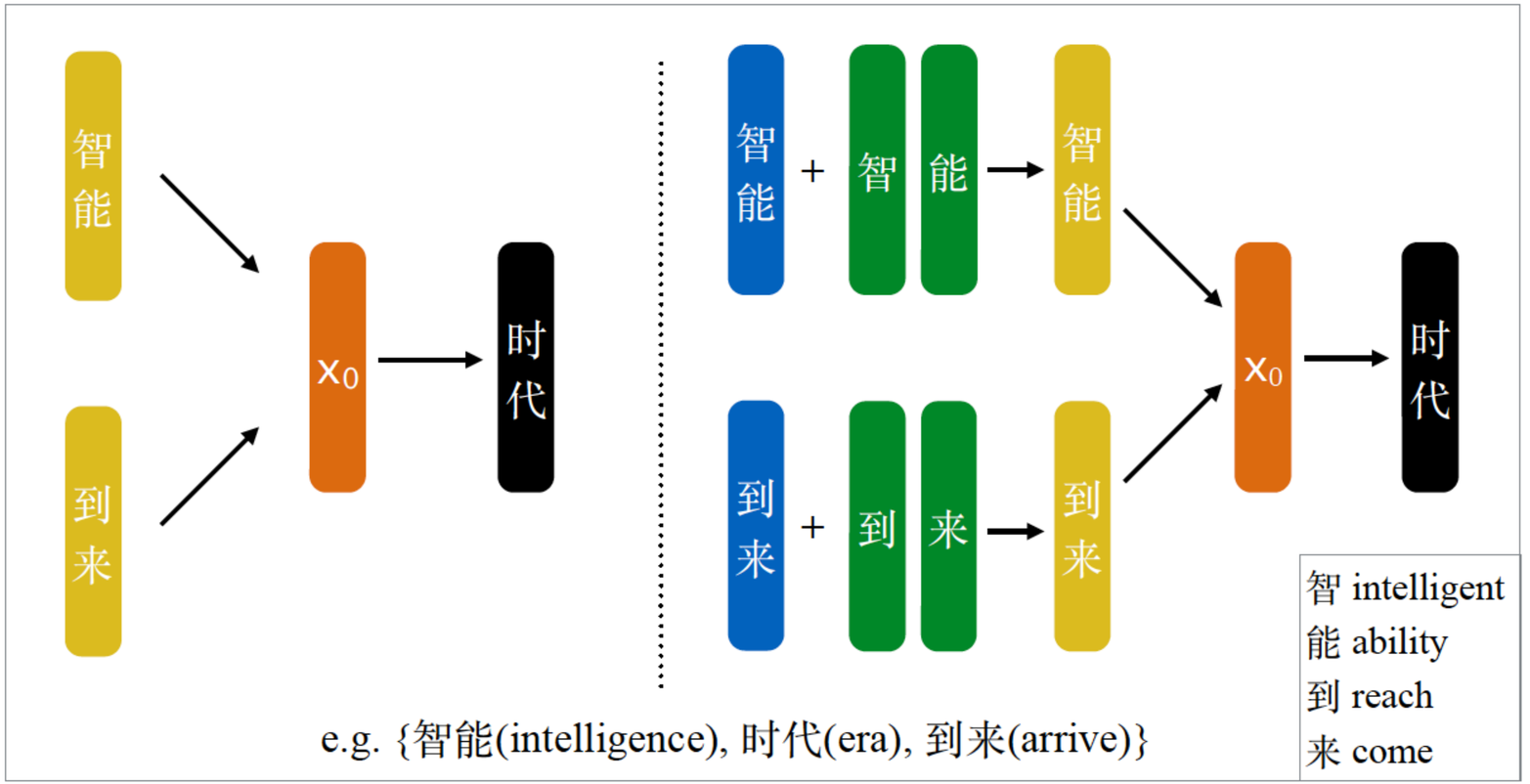
② Predict: CBOW & Skipgram (word2vec)





词嵌入概览

3．针对中文的改进：



Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and word embeddings. In IJCAI. pages 1236–1242.

类推法概览

Word analogy

类推法概览

- 引入：
  - Mikolov (2013) Linguistic regularities in continuous space word representations

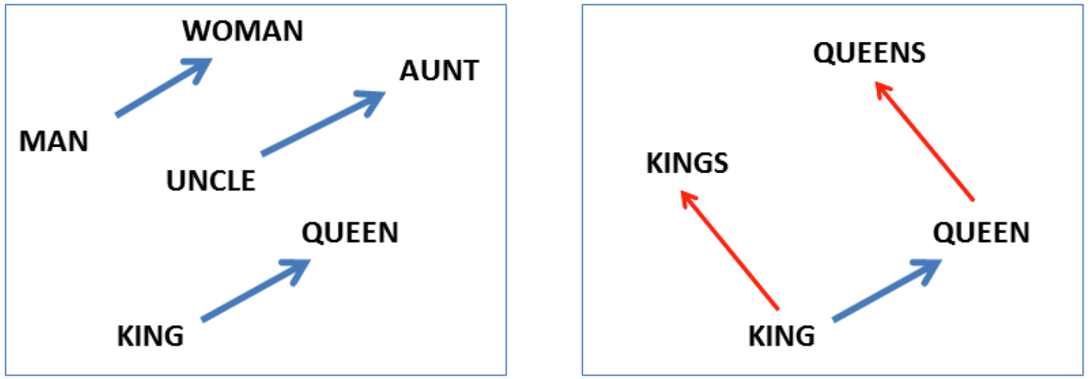


Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

类推法概览

- 类推测试的日常表述：
  - word1之于word2，相当于word3之于什么？（正解：word4）
    - 例如：东京之于日本，相当于北京之于什么？（正解：中国）

- 类推测试的日常表述：
  - word1之于word2，相当于word3之于什么？（正解：word4）
    - 例如：东京之于日本，相当于北京之于什么？（正解：中国）
- 类推测试的形式表述：

① 词汇对应关系与关系词对

$$f: V_1 \rightarrow V_2$$
$$w_1 \mapsto w_2$$

- f —— 某种对应关系，即一个映射、变换
- V_1 —— 这种语言关系中所有基式的集合
- V_2 —— 这种语言关系中所有变式的集合
- 一个关系词对就是由一个基式、一个变式及隐藏在它们背后的对应法则组成的。

eg.

$$f: V_1 \rightarrow V_2$$
$$\text{eat} \mapsto \text{ate}$$

- f —— 动词现在时-动词过去时
- V_1 —— 动词现在时的集合
- V_2 —— 动词过去时的集合

- 一个关系词对就是由一个基式、一个变式及隐藏在它们背后的对应法则组成的。

② 类推问题

- 每个类推问题会涉及两个关系词对，一共四个词：

$$f(w_a) = w_b$$
$$f(w_c) = w_d$$
$$w_a, w_c \in V_1 \quad , \quad w_b, w_d \in V_2$$

- 类推测试为了测试这种“关系”（f），会运用类比推理的逻辑来提问：

$$f(w_a) = w_b$$
$$\Rightarrow f(w_c) = ?$$
$$w_a, w_c \in V_1 \quad , \quad w_b, ? \in V_2$$

eg.

- 如果这四个词语是：w_a=eat，w_b=ate，w_c=look，w_d =looked；那么这次所测试的语法“关系”（f）就应是动词现在时和过去式之间的词法关系。

$$f(eat) = ate$$
$$\Rightarrow f(look) = ?$$
$$eat, ate \in V_1 \quad , \quad look, ? \in V_2$$

- 类推测试为了测试这种“关系”（f），会运用类比推理的逻辑来提问：

$$f(w_a) = w_b$$
$$\Rightarrow f(w_c) = ?$$
$$w_a, w_c \in V_1 \quad , \quad w_b, ? \in V_2$$

$$f(eat) = ate$$
$$\Rightarrow f(look) = ?$$
$$eat, ate \in V_1 \quad , \quad look, ? \in V_2$$

③ 用类推问题测试词向量对语言信息的捕获情况：

- 用词向量去解决一组类推问题，回答的正确率越高，则说明词向量对语言关系f的捕获能力越强。

$$\arg \max_{w'_d \in V_2} \left(sim\left(v(w'_d), v(w_c) - v(w_a) + v(w_b)\right)\right) \quad (4-5)$$

类推法概览

- 报告人的想法：

  ① 类推评测法的局限：类推测试法只能检测那些可以被设计为类推实验的linguistic regularities，而面对那些不能的（比如汉语的句法信息）就技穷了。

  ② 像类推评测法这样的词向量的内部评测方法是否有必要?

  - 有必要：有利于使词向量成为可预料性、可解释性强的一种数据表示，从而有利于增强基于机器学习在自然语言处理任务上的可解释性。
  - 没必要：从工程任务上看，从语言实际任务出发的外部评测更方便和直接。

- 注：
  - 内部评测方法：从语言本身的特性出发，侧重展示词向量是否学习到了这些linguistic regularities。如word analogy，word similarity。
  - 外部评测方法：从实际的任务出发，以实际的任务效果来反映词向量的训练好坏。通过准确率、召回率、F1指数、相关系统等指标来衡量整个系统的效果。

类推法概览

2. 作为方法的类推

① Evaluate word embeddings

② Inducing morphological transformations

③ Detecting semantic relations

④ Translating unknown words

类推法概览

2. 作为方法的类推

① Evaluate word embeddings

② Inducing morphological transformations

- 引出形态转换

③ Detecting semantic relations

- 检测语义关系

④ Translating unknown words

- 翻译未登录词

本文的类推实验设计

Experiments

类推测试集的设计

- Mikolov et al. 2013发布词类比关系数据集
  - 19,544个类比问题
    - 8869语义类比(5类)
    - 10675句法类比(9类)

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

类推测试集的设计

- CA8
  - including 17813 questions.
  - set a limit of 50 word pairs at most for each relation.
  - 1852 unique Chinese word pairs

Benchmark	Category	Type	#questions	#words	Relation
CA_translated	Semantic	Capital	506	46	capital-country
		State	175	54	city-province
		Family	272	34	family members
CA8	Morphological	Reduplication A	2554	344	A-A, A-yi-A, A-lái-A-qù
		Reduplication AB	2535	423	A-A-B-B, A-lǐ-A-B, A-B-A-B
		Semi-prefix	2553	656	21 semi-prefixes: 大, 小, 老, 第, 亚, etc.
		Semi-suffix	2535	727	41 semi-suffixes: 者, 式, 主义, 性, etc.
	Semantic	Geography	3192	305	country-capital, country-currency, province-abbreviation, province-capital, province-dramma, etc.
		History	1465	177	dynasty-emperor, dynasty-capital, title-emperor, celebrity-country
		Nature	1370	452	number, time, animal, plant, body, physics, weather, reverse, color, etc.
		People	1609	259	finding-scientist, work-writer, family members, etc.

汉语中的词法类推

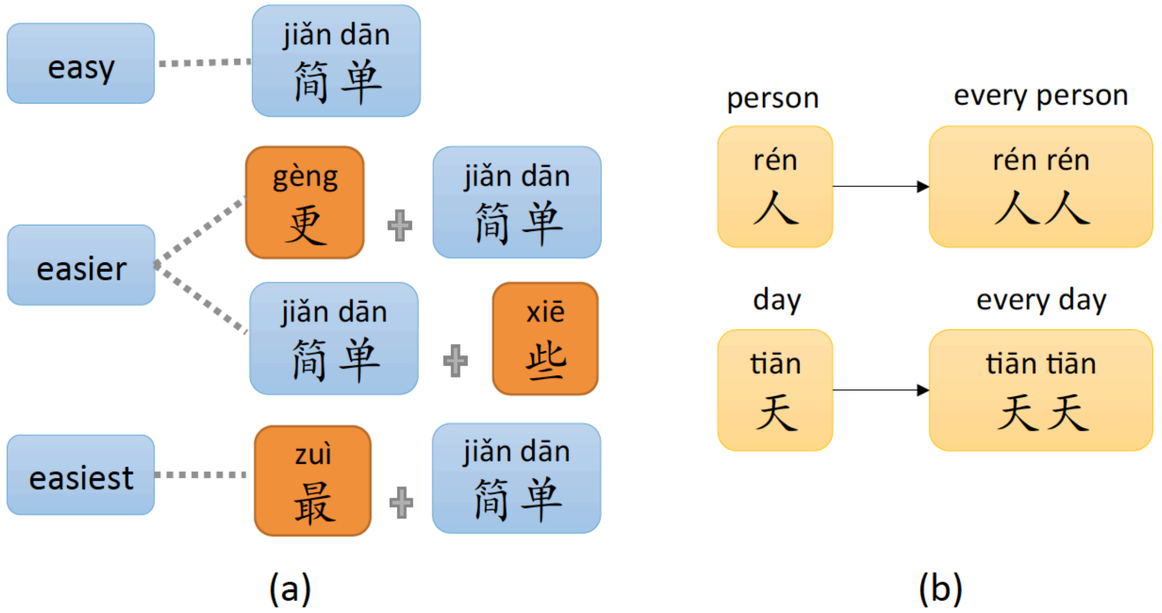


Figure 1: Examples of Chinese lexical knowledge: (a) function words (in orange boxes) are used to indicate the comparative and superlative degrees; (b) reduplication yields the meaning of "*every*".

汉语中的词法类推

1. 重叠
2. 加缀

Morphological Questions: Reduplication

Morphological Questions: Reduplication				
Category	Sub-category	POS	Morphological Function	Example
A	AA	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)
		Noun	Yield many / much meaning	山水 (mountain and river) → 山山水水 (many mountains and rivers)

	qu		repeatedly	
AB	AABB	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)
	ABAB	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)

2.加缀

Morphological Questions: Semi-affixation		
Category	Semi-affix	Example
	们	我 (I) → 我们 (we)
	里	这 (here) → 这里 (here)
	些	这 (this) → 这些 (these)

Morphological Questions: Semi-affixation		
Category	Semi-affix	Example
Semi-prefix	第	一 (one) → 第一 (first)
	初	一 (one) → 初一 (the first day of a lunar month)
	十	一 (one) → 十一 (eleven)
	周	一 (one) → 周一 (Monday)
	星期	一 (one) → 星期一 (Monday)
	老	虎 (tiger) → 老虎 (tiger)
	小	草 (grass) → 小草 (grass)
	大	海 (sea) → 大海 (large sea)
	半	导体 (conductor) → 半导体 (semiconductor)
	单	细胞 (cell) → 单细胞 (unicell)
	超	链接 (link) → 超链接 (hyperlink)
	次	大陆 (continent) → 次大陆 (subcontinent)
	非	常规 (conventional) → 非常规 (unconventional)
	每	次 (time) → 每次 (every time)
	全	明星 (star) → 全明星 (all star)
	伪	君子 (gentlemen) → 伪君子 (hypocrites)
	亚	热带 (tropical zone) → 亚热带 (sub-tropical zone)
	洋	酒 (wine) → 洋酒 (foreign wine)
	总	比分 (score) → 总比分 (total score)
	反	物质 (matter) → 反常规 (antimatter)
	副	总统 (president) → 副总统 (vice president)
Semi-suffix	们	我 (I) → 我们 (we)
	里	这 (here) → 这里 (here)
	些	这 (this) → 这些 (these)
	样	这 (this) → 这样 (such)
	个	这 (this) → 这个 (this one)
	边	这 (this) → 这边 (here)
	种	这 (this) → 这种 (this kind)
	次	这 (this) → 这次 (this time)
	儿	这 (this) → 这儿 (here)
	部	东 (east) → 东部 (east)
	中	心 (heart) → 心中 (in the heart)
	上	山 (mountain) → 山上 (on the mountain)
	面	前 (front) → 前面 (in the front)
	者	强 (strong) → 强者 (the strong one)
	家	科学 (science) → 科学家 (scientist)
	子	胖 (fat) → 胖子 (a fat man)
	头	木 (wood) → 木头 (wood)
	工	木 (wood) → 木工 (carpentry)
	匠	木 (wood) → 木匠 (carpenter)
	星	笑 (laugh) → 笑星 (comedian)
	手	老 (old) → 老手 (old hand)
	主义	乐观 (optimistic) → 乐观主义 (optimism)
	鬼	吝啬 (stingy) → 吝啬鬼 (miser)
	式	中 (Chinese) → 中式 (Chinese style)
	队	考古 (archaeology) → 考古队 (archaeological team)
	色	黄 (yellow) → 黄色 (the yellow color)
	学	地质 (geology) → 地质学 (discipline of geology)

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

汉语中的词法类推

- 报告人的想法：
  - 利用词的内部曲折方式构建词重叠的类推测试集，基本思路可行。
  - 但是：
    A．重叠测试集中所列的“语言单位”，诸如“姥”、“坛/罐”、“想一想”、“运动运动” 等，这些是否应该处理为词，需要商榷。并且，切词工具能否切出这些单位，也是问题。
    - 因此，报告人建议：
      ① 在重叠上，应区分音重叠、重叠式合成词和词重叠。
      ② 由于本文的研究对象是词向量，考察的应主要是词和词的语法变体，而像“姥-姥姥”“坛/罐-坛坛罐罐”这样的音重叠或重叠式合成词，因为它们的基式“姥”“坛/罐”等本身不是词，本身研究价值也不大，因此在这里应不作研究。
      ③ 而像“想一想”、“运动运动”这样的词组，比较容易通过加大切词工具的颗粒度来切出，并且有一定研究价值，因此可以考虑纳入研究。

汉语中的词法类推

- 报告人的想法：
  - 利用词的内部曲折方式构建词重叠的类推测试集，基本思路可行。
  - 但是：
    B．词法类推，特别是重叠类推，是很确切的事，根据语言学知识（规则或词库，如GKB）即可得到，那么词向量捕获词法信息在计算语言学学科中是否具有实用价值?

汉语中的词法类推

- 报告人的想法：

  C．词法类推测试集，特别是加缀测试集，设计上有些不足。许多词对在语义上的类推关系已经大于其词法上的类推关系了，最终不知道到底是在检验语义信息还是检验语法信息；并且，尚不论很多缀，实际是否是缀在语言学界还有很大争议。

  - 因此，报告人建议：

    ① 重新审视加缀测试集，剔除不恰当的项目，保留“子”“儿”“头”等经典项目。

    ② 把从加缀测试集剔出来的那些项目另立名目，加入语义类推测试集中去。

汉语中的语义类推

汉语中的语义类推

1.地理

2.历史

3.自然

4.人

Semantic Questions			

Semantic Questions		
Category	Sub-category	Example
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)

本文的类推实验设计

- 类推测试集
  - CA8
  - CA-translated
- 训练模型
  - SGNS (skip-gram model with negative sampling)
  - PPMI (Positive Pointwise Mutual Information)

Window (dynamic)	Iteration	Dimension	Sub-sampling	Low-frequency threshold	Context distribution smoothing	Negative (SGNS/PPMI)	Vector offset
5	5	300	1e-5	50	0.75	5/1	3COSMUL

- 输入层
  - word
  - word+character

本文的类推实验设计

- 训练语料

Corpus	Size	#tokens	$\|V\|$	Description
Wikipedia	1.3G	223M	2129K	Wikipedia data obtained from https://dumps.wikimedia.org/
Baidubaike	4.1G	745M	5422K	Chinese wikipedia data from https://baike.baidu.com/
People's Daily News	3.9G	668M	1664K	News data from People's Daily (1946-2017) http://data.people.com.cn/
Sogou news	3.7G	649M	1226K	News data provided by Sogou Labs http://www.sogou.com/labs/
Zhihu QA	2.1G	384M	1117K	Chinese QA data from https://www.zhihu.com/, including 32137 questions and 3239114 answers
Combination	14.8G	2668M	8175K	We build this corpus by combining the above corpora

实验结果

Results

实验结果

- 模型对比

		CA_translated			CA8								
		Cap.	Sta.	Fam.	A	AB	Pre.	Suf.	**Mor.**	Geo.	His.	Nat.	Peo.	**Sem.**
SGNS	word	.706	.966	.603	.117	.162	.181	.389	.222	.414	.345	.236	.223	.327
	word+ngram	.715	**.977**	.640	.143	.184	.197	.429	.250	.449	.308	.276	.310	.368
	word+char	.676	.966	.548	**.358**	**.540**	**.326**	**.612**	**.455**	.468	.226	.296	.305	.368
PPMI	word	.925	.920	.548	.103	.139	.138	.464	.226	.627	.501	.300	.515	.522
	word+ngram	**.943**	.960	**.658**	.102	.129	.168	.456	.230	**.680**	**.535**	**.371**	**.626**	**.586**
	word+char	.913	.886	.614	.106	.190	.173	.505	.260	.638	.502	.288	.515	.524

- SGNS 在捕获词法信息方面表现优越
- PPMI 在捕获语义信息方面表现优越
- 与英语的情况一致
- 本文猜测的原因：

  ① 词法信息更加依赖高频词，而SGNS擅长捕获高频词的信息。

  ② 语义信息更加依赖低频词，而PPMI擅长捕获低频词的信息。

实验结果

- 输入层对比

		CA_translated			CA8								
		Cap.	Sta.	Fam.	A	AB	Pre.	Suf.	**Mor.**	Geo.	His.	Nat.	Peo.	**Sem.**
SGNS	word	.706	.966	.603	.117	.162	.181	.389	.222	.414	.345	.236	.223	.327
	word+ngram	.715	**.977**	.640	.143	.184	.197	.429	.250	.449	.308	.276	.310	.368
	word+char	.676	.966	.548	**.358**	**.540**	**.326**	**.612**	**.455**	.468	.226	.296	.305	.368
PPMI	word	.925	.920	.548	.103	.139	.138	.464	.226	.627	.501	.300	.515	.522
	word+ngram	**.943**	.960	**.658**	.102	.129	.168	.456	.230	**.680**	**.535**	**.371**	**.626**	**.586**
	word+char	.913	.886	.614	.106	.190	.173	.505	.260	.638	.502	.288	.515	.524

- 在词法信息捕获方面，引入字向量的SGNS明显好于其他
- 在语义信息捕获方面，引入ngram的PPMI明显好于其他

实验结果

- 训练语料对比

	CA_translated			CA8								
	Cap.	Sta.	Fam.	A	AB	Pre.	Suf.	**Mor.**	Geo.	His.	Nat.	Peo.	**Sem.**
Wikipedia 1.2G	.597	.771	.360	.029	.018	.152	.266	.180	.339	.125	.147	.079	.236
Baidubaike 4.3G	.706	.966	.603	.117	.162	.181	**.389**	.222	.414	**.345**	.236	**.223**	.327
People's Daily 4.2G	**.925**	**.989**	.547	.140	.158	**.213**	.355	**.226**	**.694**	.019	.206	.157	**.455**
Sogou News 4.0G	.619	.966	.496	.057	.075	.131	.176	.115	.432	.067	.150	.145	.302
Zhihu QA 2.2G	.277	.491	**.625**	**.175**	**.199**	.134	.251	.189	.146	.147	**.250**	.189	.181
Combination 15.9G	**.872**	**.994**	**.710**	**.223**	**.300**	**.234**	**.518**	**.321**	**.662**	**.293**	**.310**	**.307**	**.467**

  - 随着语料规模的增加，正确率提高
  - 语料的性质也影响正确率，例如：
    - 新闻性质的语料在语义问题（尤其是地理类问题）上表现最好。
    - 问答性质的语料在重叠问题上表现最好。
      - 本文推测原因：问答社区包含更多口语，而汉语的重叠多出现在口语中。

ANALOGICAL REASONING ON
CHINESE MORPHOLOGICAL AND SEMANTIC RELATIONS

汉语的词向量与类推测试

北京大学 中文信息处理 唐乾桐

1.问题描述
2.词嵌入概览
3.类推法概览
4.本文的类推实验设计
5.实验结果