WORD CLASSIFICATION FROM THE PERSPECTIVE OF WORD EMBEDDING

从词向量的视角审视词类问题

北京大学 中文信息处理 唐乾桐

词向量概览

Word Embedding

词向量概览

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

词向量概览

2. 获得词向量的常见方法:

① Count (Sparse)

① PPMI

② Predict (Dense)

① C&W

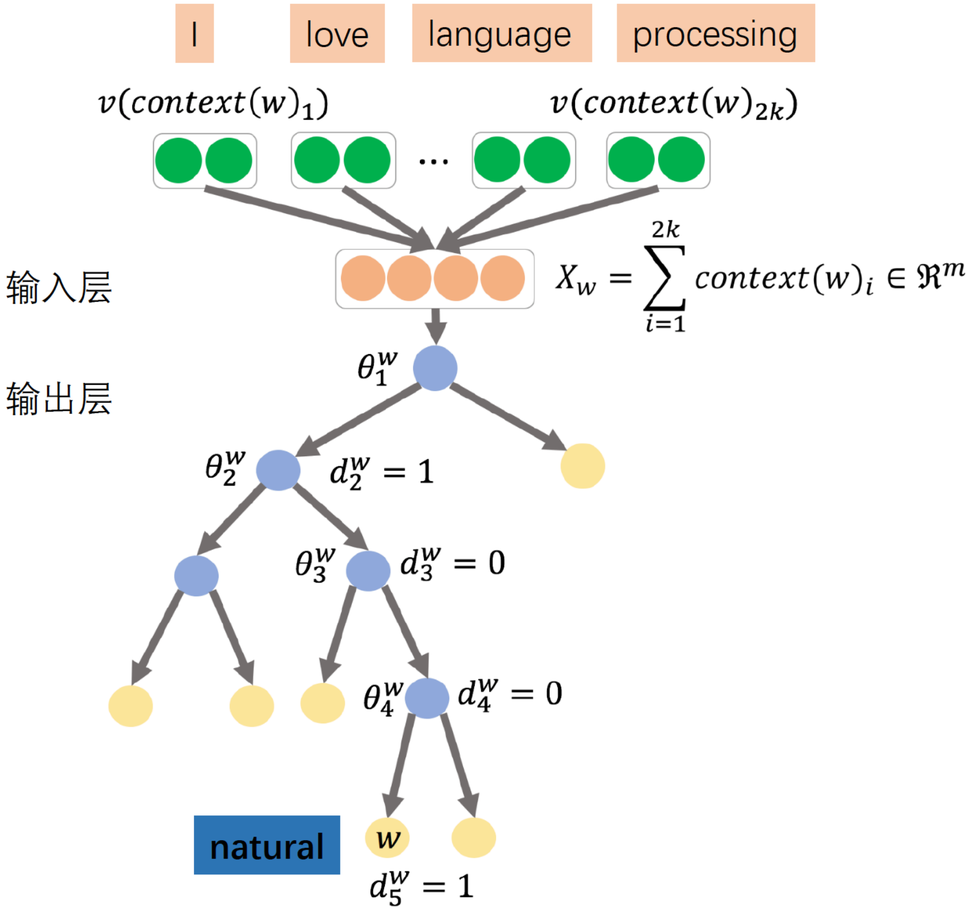
② CBOW & Skipgram (word2vec)

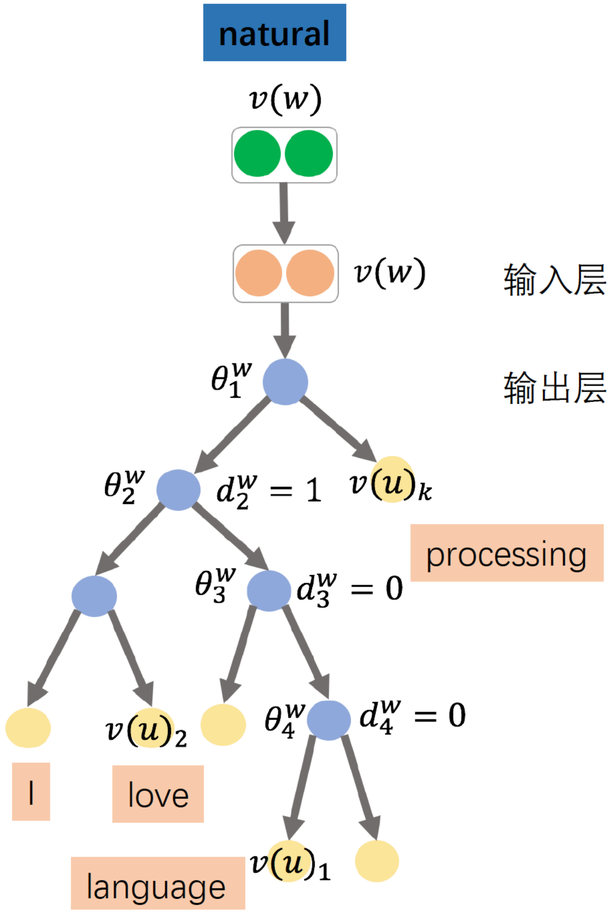
③ Transformers (Bert)

相关文献：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.

词向量概览

② Predict: CBOW & Skipgram (word2vec)





相关文献：2018年秋季组会报告

词类体系

Word class systems

词类体系

各家的词类体系概览

出处	作者	年份	分类数量
《马氏文通》	马建忠	1898	九类
《新著国文语法》	黎锦熙	1924	九类
《中国文法要略》	吕叔湘	1941	七类
《中国现代文法》	王力	1954	九类
《现代汉语八百词》	吕叔湘	1980	十三类
《语法讲义》	朱德熙	1982	十七类
《中学教学语法系统提要》	——	1984	十二类
《现代汉语》	刑福义	1991	十一类
《现代汉语》	北大现代汉语教研室	1993	十五类
《现代汉语》	胡裕树	1995	十三类
《现代汉语》	黄伯荣、廖序东	1997	十四类
《现代汉语》	沈阳、郭锐等	2014	二十类

词类体系

各家的词类体系概览

出处	作者	年份	分类数量
《马氏文通》	马建忠	1898	九类
《新著国文语法》	黎锦熙	1924	九类
《中国文法要略》	吕叔湘	1941	七类
《中国现代文法》	王力	1954	九类
《现代汉语八百词》	吕叔湘	1980	十三类
《语法讲义》	朱德熙	1982	十七类
《中学教学语法系统提要》	——	1984	十二类
《现代汉语》	刑福义	1991	十一类
《现代汉语》	北大现代汉语教研室	1993	十五类
《现代汉语》	胡裕树	1995	十三类
《现代汉语》	黄伯荣、廖序东	1997	十四类
《现代汉语》	沈阳、郭锐等	2014	二十类

汉语的词类到底有多少个？分别是哪几个？各家观点极少有完全相同的。这是汉语词类体系研究走向精密化、科学化必须要解决的问题。

汉语词语的语法信息处理，相较于英语等形态丰富的语言，需要更多地依赖词语的功能分布信息；而词向量基于词的上下文来对词进行向量建模的方式，正是一种分布表示(Distributional representation) 方法。因而，借助词向量来表示汉语词语的语法信息，尤其是词类信息，是非常值得考虑的尝试；在词向量的背景下，探讨一个统一的、颗粒度可调、面向不同应用的词类方案，也是非常值得尝试的课题。

用词向量归词类，是否可行?

Word Embeddings and the Grammatical Information

用词向量归词类，是否可行?

【大致思路】用聚类算法直接聚类。这里词向量选择CBOW，聚类算法选择混合高斯。

【结果】

```
cluster_table = Table().with_column('词语', vocab, '类别', label, '向量'
clus          ort('类别')
```

聚类结果

正确词类

词语	类别	向量	词类
较	0	[-2.11442839e-02 -4.51438352e-02 -7.05480501e-02 7.8914 ...	['p' 'd']
价格	0	[0.03092024 -0.01391196 0.03561808 -0.01205909 0.0503 ...	['n']
增长	0	[0.06036939 -0.0199586 0.06783856 -0.08107521 -0.1002 ...	['v']
影响	0	[-0.00761042 -0.04233571 0.03150782 0.01573941 0.1149 ...	['v']
数据	0	[0.00955876 -0.0377751 -0.11274419 -0.04818853 0.0009 ...	['n']
股	0	[-1.32712405e-02 -1.30173951e-01 -4.63096388e-02 6.3407 ...	['q' 'q' 'n' 'n']
下降	0	[1.50116095e-02 -7.32336268e-02 2.67775282e-02 3.8722 ...	['v']
表现	0	[-1.98896471e-02 1.53018460e-02 -6.67595789e-02 -1.0716 ...	['v']
调整	0	[-7.05896839e-02 -7.98866302e-02 -3.61891203e-02 5.9538 ...	['v']
上半年	0	[1.17807938e-02 -3.91006507e-02 -3.97105515e-02 -7.1653 ...	['t']

【结论】直接聚类不可取。通过聚类程序自动找到的类别，语法上的相似性较弱。词向量捕获的语言信息中，除了语法信息之外，还含有大量语义信息。怎么把其中的语法信息单独抽离出来？词向量做词的语法聚类，是否可行？这是本研究最大的一个重难点。

用词向量归词类，是否可行?

思路需要更新

用词向量归词类，是否可行?

基于原型的现代范畴化理论认为:
(1)范畴不一定能用一组充分必要特征/条件来下定义，在区别一个范畴时，没有一个属性是必要的;
(2)实体的范畴化是建立在好的、清楚的样本(exemplar)的基础之上的，然后将其他实例根据它们跟这些好的、清楚的样本在某些/一组属性上的相似性而归入该范畴;
(3)这些好的、清楚的样本就是典型(即原型)，它们是非典型事例范畴化的参照点。
这种根据与典型事例类比而得出的范畴就是原型范畴(prototype based category)。

相关文献：袁毓林. **词类是原型范畴**. 1995

用词向量归词类，是否可行?

【大致思路】根据原型范畴理论:

① 选取各大主流的现代汉语词类体系都普遍承认的几个词类（如动词、名词、形容词等）

② 找到这些词类的好的、清楚的样本（如动词的“打”“吃”等，名词的“桌子”“太阳”等）及其向量表示，作为典型（原型）。它们是非典型词语范畴化的参照点。

③ 计算其他词的向量表示与这些典型之间的余弦距离（相似性），并以此为依据进行颗粒度可控的归类。

- 【困难】仍然是那个问题：词向量捕获的语言信息中，除了语法信息之外，还含有大量语义信息。余弦距离是否能准确抽离语法信息?

相关文献：袁毓林. **词类是原型范畴**. 1995

用词向量归词类，是否可行?

- 【测试】引入类推测试进行测试：

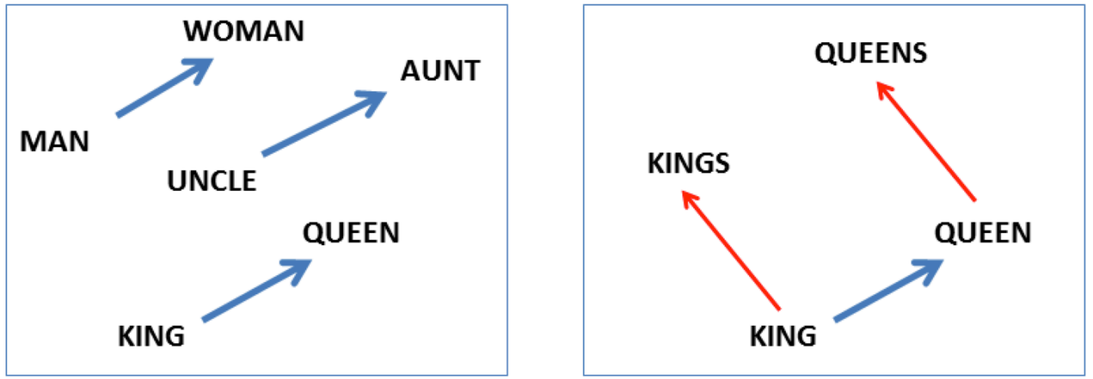


Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

相关文献：Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013c: 746-751.

- 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

相关文献：2018年秋季组会报告

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。

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

相关文献：2018年秋季组会报告

用词向量归词类，是否可行?

- 引入类推测试进行测试：

③ 用类推问题测试词向量对语法信息的捕获情况：

- 用词向量去解决一组测试语法关系的类推问题，回答的正确率越高，则说明词向量对语法关系f的捕获能力越强。

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

相关文献：2018年秋季组会报告

用词向量归词类，是否可行?

- 引入类推测试进行测试：

  ① 词汇对应关系与关系词对

  ② 类推问题

  ③ 用类推问题测试词向量对语法信息的捕获情况

用词向量归词类，是否可行?

- 引入类推测试进行测试：

  ① 词汇对应关系与关系词对

  ② 类推问题

  ③ 用类推问题测试词向量对语法信息的捕获情况

【困难】汉语不同于英语等综合语，它没有丰富的形态变化，因此就缺少可用作语法信息类推测试的关系词对，给类推测试的设计带来了困难。

【解决】为了克服这个困难，首先从控制变量的基本原则出发，选取语义基本相同、但语法功能明显不同的若干词对（如：“突然—忽然”，“刚才—刚刚”等），然后按它们所反映的具体词类变化关系对它们进行分类，再按类别来生成类推问题。

用词向量归词类，是否可行?

- 引入类推测试进行测试：

为了选取语义基本相同、但语法功能明显不同的词对，首先集成《现代汉语语法信息词典》（简称GKB）和《哈工大社会计算与信息检索研究中心同义词词林扩展版》两个资源。对于每个词语，从前者提取其词类信息，从后者提取其同义词，再辅以人工校对。

然而，正如朱学锋等（2004）指出，“如何克服不同知识库之间的’缝隙’将成为集成不同语言数据资源时无可回避的普遍问题”，当综合利用GKB和《哈工大社会计算与信息检索研究中心同义词词林扩展版》这两个资源库的时候，也发现了它们之间存在着的“缝隙”或“鸿沟”(gap)，其中最大的一个“缝隙”就是两个资源库所收录的词语不完全一致，有些词在一个资源库里收录了，但在另一个资源库里没有收录；或者虽然都收录了，但只是词形上的一致。比如，“大方”在GKB中只收录了形容词（慷慨义），而在《哈工大社会计算与信息检索研究中心同义词词林扩展版》中还收录了名词（表示某领域内的专家）。为了解决这个问题，我以两个资源库中都收录了的双音节词为基础，辅以逐条的人工校对，以排除只是词形一致的情况。

相关文献：朱学锋,张化瑞,段慧明,俞士汶. 《汉语高频词语法信息词典》的研制[J].语言文字应用,2004(03):98-104.

用词向量归词类，是否可行?

- 引入类推测试进行测试：

  我提取了3,696对双音节词词对，依据词类组合的不同将他们分为574个小类，依据这些小类设计了一共201,988个测试问题（尚需人工筛查）。

序号	类别	词对数量	问题数量	问题举例
1	副词:形容词	302	35989	忽然:突然⇒牢牢:（?）
2	动词:形容词	207	30236	缺乏:贫乏⇒怠慢:（?）
3	形容词:名词	131	13012	友好:友谊⇒淫秽:（?）
4	区别词:形容词	90	2256	优等:优质⇒首要:（?）
5	动词:副词	87	3296	赶紧:从速⇒应当:（?）
6	动词:名词	59	2675	琢磨:思想⇒企图:（?）
7	时间词:副词	44	729	刚才:刚刚⇒平时:（?）
8	[区别词/副词][4]:形容词	39	578	快速:迅速⇒日常:（?）
9	副词:连词	39	930	随后:而后⇒一旦:（?）
10	形容词:状态词	31	652	高大:巍然⇒白净:（?）
…	…	…	…	…
汇总		3696	201988	

相关文献：俞士汶,朱学锋,王惠,张芸芸. 现代汉语语法信息词典规格说明书[J].中文信息学报,1996(02):1-22.

词向量捕获语法信息

- 引入类推测试进行测试：

  测试结果：

  - CBOW与Skim-gram相比，每行正确率高者用黑体标出。
  - 从所有答对的案例中随机抽取1例，其格式为：w1:w2 => w3:w4。其中w1、w2、w3为输入，为w4为输出，后表同。

序号	类型	有效问题数量	正确率(%)[4]		答对举例[5]
			CBOW	SKIM	
1	名词:时间词	157	**15.28**	6.37	青春:春季⇒尾声:(暮年)
2	名词:处所词	428	**16.12**	13.79	大陆:陆上⇒天空:(上空)
3	名词:方位词	195	**25.64**	10.77	边缘:旁边⇒末尾:(后面)
4	名词:动词	2387	**14.16**	5.95	耻辱:污辱⇒反响:(共鸣)
5	名词:形容词	8771	**26.20**	8.90	朋友:友好⇒细节:(细致)
6	时间词:方位词	337	**32.93**	30.27	先前:以前⇒往后:(以后)
7	时间词:副词	729	**26.20**	19.20	平日:平素⇒后来:(随后)
8	时间词:连词	72	**34.72**	15.28	往后:而后⇒往后:(然后)
9	处所词:名词	397	**15.86**	3.53	乡下:乡村⇒路上:(道路)
10	处所词:方位词	17	**29.41**	0.00	外地:外边⇒幕后:(底下)
11	处所词:副词	104	**35.57**	14.42	心中:满心⇒心里:(暗自)
12	方位词:名词	379	**23.48**	14.25	上面:上级⇒下面:(下级)
13	方位词:时间词	179	**32.96**	29.61	之前:事前⇒之后:(事后)
14	方位词:处所词	17	**17.64**	5.88	前后:近处⇒外边:(远处)
15	方位词:副词	206	**37.86**	15.05	之前:事先⇒以来:(至今)
16	数词:形容词	101	**28.71**	2.97	许多:浩大⇒众多:(庞大)
17	动词:名词	2675	**15.92**	7.21	污辱:耻辱⇒束缚:(枷锁)
18	动词:形容词	15137	**23.28**	9.42	怠慢:不周⇒熟悉:(常见)
19	动词:副词	3296	**21.63**	17.60	邻近:就近⇒无疑:(必定)
20	动词:连词	26	**42.30**	34.62	加以:况且⇒相反:(但是)
21	形容词:名词	13012	**13.51**	3.67	友好:朋友⇒贤惠:(老婆)
22	形容词:数词	67	**53.73**	52.24	浩大:许多⇒有限:(一些)
23	形容词:动词	30236	**17.01**	8.99	崎岖:起伏⇒贴心:(体贴)
24	形容词:状态词	652	**2.60**	0.46	细小:细高⇒白净:(雪白)
25	形容词:区别词	5525	**2.80**	1.63	适中:中等⇒特殊:(特定)
26	形容词:副词	55589	**15.00**	6.21	频繁:屡屡⇒突然:(忽然)
27	形容词:[区别词/副词]	1371	**3.71**	1.31	迅速:快速⇒随意:(任意)
28	形容词:[副词/区别词]	872	**1.60**	0.34	悠久:长期⇒常见:(通常)

词向量捕获语法信息

- 引入类推测试进行测试：

  测试结果：

  - CBOW与Skim-gram相比，每行正确率高者用黑体标出。
  - 从所有答对的案例中随机抽取1例，其格式为：w1:w2 => w3:w4。其中w1、w2、w3为输入，为w4为输出，后表同。

29	形容词:[副词/动词]	650	0.15	**0.77**	频繁:反复⇒努力:(尽力)
31	形容词:[数词/副词]	30	**36.66**	33.33	豁达:大量⇒有限:(少量)
32	状态词:形容词	382	**25.91**	8.90	巍然:高大⇒细高:(瘦小)
33	状态词:副词	63	**36.50**	26.98	滚圆:团团⇒频频:(屡屡)
34	区别词:名词	73	**12.32**	4.11	五金:金属⇒亲生:(骨肉)
35	区别词:时间词	28	**46.42**	39.29	现行:现在⇒现行:(如今)
36	区别词:方位词	27	**66.66**	59.26	中等:当中⇒中档:(之中)
37	区别词:动词	140	**10.00**	0.00	次要:辅助⇒便民:(服务)
38	区别词:形容词	2256	**29.52**	8.16	劣质:低劣⇒首要:(重要)
39	区别词:副词	108	**24.07**	7.41	慢性:徐徐⇒首任:(缓缓)
41	副词:名词	155	**15.48**	2.58	多方:多边⇒秉公:(准则)
42	副词:时间词	690	**14.92**	12.32	平素:平时⇒新近:(近期)
43	副词:方位词	264	**25.37**	11.74	自古:以来⇒事先:(之前)
44	副词:动词	3953	**8.80**	4.98	就近:邻近⇒蓄意:(诽谤)
45	副词:形容词	35989	**30.29**	8.51	紧紧:严密⇒飞速:(迅速)
46	副词:状态词	198	**2.53**	**2.53**	不住:连连⇒不断:(频频)
47	副词:区别词	240	3.33	**10.83**	一再:高频⇒初次:(低频)
48	副词:连词	930	31.29	**32.80**	随后:而后⇒随后:(然后)
49	副词:[形容词/动词]	134	**14.92**	13.43	只顾:小心⇒欣然:(高兴)
51	副词:[名词/动词]	110	**4.54**	0.00	决计:决定⇒决意:(打算)
52	副词:[时间词/时间词]	170	**2.94**	0.00	立刻:当时⇒当场:(当年)
53	介词:副词	20	**5.00**	0.00	趁着:乘势⇒趁着:(趁势)
54	连词:时间词	24	**29.16**	25.00	而后:今后⇒而后:(未来)
55	连词:动词	26	**34.61**	15.38	况且:加以⇒何况:(予以)
56	[区别词/副词]:动词	26	**46.15**	15.38	临场:出席⇒额外:(赠送)
57	[区别词/副词]:形容词	578	**35.81**	8.82	快速:敏捷⇒急剧:(暴躁)
58	[副词/区别词]:形容词	306	**22.87**	7.19	长期:遥远⇒通常:(寻常)
59	[副词/动词]:形容词	252	**36.50**	20.24	当真:认真⇒尽力:(积极)
60	[副词/动词]:[副词 名词]	17	**17.65**	**17.65**	着意:苦心⇒到底:(究竟)
61	[名词/动词]:副词	62	**12.90**	6.45	开始:从头⇒可能:(恐怕)
62	[形容词/名词]:副词	16	**12.50**	0.00	实际:其实⇒理想:(也许)
63	[名词/副词]:形容词	38	**26.31**	15.79	大概:简易⇒决心:(坚决)
64	[副词/名词]:[副词/动词]	17	**17.65**	**17.65**	全力:尽力⇒究竟:(到底)
65	[时间词/时间词]:副词	39	**56.41**	41.03	当时:当场⇒当年:(一举)
…	…	…	…	…	…
汇总		**201988**	**19.15**	**7.67**	

用词向量归词类，是否可行?

- 引入类推测试进行测试：

【结论】

- 至少CBOW与Skim-gram在词类信息的捕获上，表现比较差。直接用词向量去给词做语法上的聚类，期望得到词的语法类（即通常所说的“词类”），效果恐怕不好。

【不足】

- 观察我们得到的词对数据集，其实发现：许多词对在语义上的类推关系已经大于其词类上的类推关系了，最终不知道到底是在检验语义信息还是检验语法信息。

用词向量归词类，是否可行?

【之前的思路】根据原型范畴理论:

① 选取各大主流的现代汉语词类体系都普遍承认的几个词类（如动词、名词、形容词等）

② 找到这些词类的好的、清楚的样本（如动词的“打”“吃”等，名词的“桌子”“太阳”等）及其向量表示，作为典型（原型，prototype）。它们是非典型词语范畴化的参照点。

③ 计算其他词的向量表示与这些典型之间的距离（相似性），并以此为依据进行颗粒度可控的归类。

【更新的思路】根据原型范畴理论:

① 选取各大主流的现代汉语词类体系都普遍承认的几个词类（如动词、名词、形容词等）

② 找到这些词类的好的、清楚的样本（如动词的“打”“吃”等，名词的“桌子”“太阳”等）及其向量表示，作为典型（原型，prototype）。它们是非典型词语范畴化的参照点。

③ 将这些prototype作为训练集，他们的词向量作为输入，他们的词类作为输出。这个输出可以是一个向量，向量的每个维对应一个代表概率的输出，prototype的输出是独热的。训练的得到的模型去预测其他词，其他词的输出不是独热的。以此为依据进行颗粒度可控的归类。

下一步

what to do next

下一步

1. 试一试比较火的预训练模型（如Bert），或许效果能提高很多。
    - 【困难】官方发布的简体中文模型是以字为粒度进行切分，没有考虑到中文分词。
    - 【思路】
        ① 寻找其他的以词为粒度的简体中文Bert项目，如Chinese-BERT-wwm。
            - 【可能的困难】因为训练过程不受自己控制，在后期需要做比较研究时，可能会遇到其他麻烦。

        ② 自己用官方发布的代码进行训练，这样可以做到全程可控。
            - 【可能的困难】受制于计算能力。

https://github.com/google-research/bert

https://github.com/ymcui/Chinese-BERT-wwm

下一步

2. 【更新的思路】根据原型范畴理论：

① 选取各大主流的现代汉语词类体系都普遍承认的几个词类（如动词、名词、形容词等）

② 找到这些词类的好的、清楚的样本（如动词的“打”“吃”等，名词的“桌子”“太阳”等）及其向量表示，作为典型（原型，prototype）。它们是非典型词语范畴化的参照点。

③ 将这些prototype作为训练集，他们的词向量作为输入，他们的词类作为输出。这个输出可以是一个向量，向量的每个维对应一个代表概率的输出，prototype的输出是独热的。训练的得到的模型去预测其他词，其他词的输出不是独热的。以此为依据进行颗粒度可控的归类。

下一步

3. 重点考察对人而言归类困难的词，看这些词会被如何归类，视情况调整下一步做法。
   - 参考刑福义《词类辩难》，搜集这样的词：
     - “最近”“最后”“末了”“早”“清早”“一清早”“原来”“一生”
     - “类似”“仿佛”“够”“留神”“是”“一定”
     - “正”“副”“绝对”“相对”“内在”“外在”“男”“女”“公”“母”“雌”“雄”“故意”“难免”“单独”“无谓”“许久”“抱歉”“抱愧”“整”“整整”
     - “继续”“开始”“恐怕”“一律”“附带”“难以”“足以”“宁可”“宁肯”“宁愿”“反而”“反倒”
     - “许多”“多”“好些”“无数”“千万”“半”“一带”
     - “一切”“任何”“所有”“全”“全体”“全部”“各位”“诸位”“整个”“凡”“凡是”“大凡”“所谓”“另外”“旁”“旁人”“人”“个人”
     - “本着”“论”“归”“临”“赶”“等”“等到”“拿”“替”
     - “然后”“至于”“果然”“万一”“比方”“一旦”“管”“别管”
     - “多”“来”“开外”“等、等等、云云、云”“不过”“连”“起见”“的”

WORD CLASSIFICATION FROM THE PERSPECTIVE OF WORD EMBEDDING

从词向量的视角审视词类问题

北京大学 中文信息处理 唐乾桐