Multi-channel Reverse Dictionary Model
多通道反向词典模型

THUNLP
岂凡超
qfc17@mails.tsinghua.edu.cn
2019/12/22

研究背景—反向词典

定义: A wide road that allows traffic to travel fast.

（普通）词典

expressway

研究背景—反向词典

定义: A wide road that allows traffic to travel fast.

（普通）词典

expressway | freeway | motorway | ……

反向词典

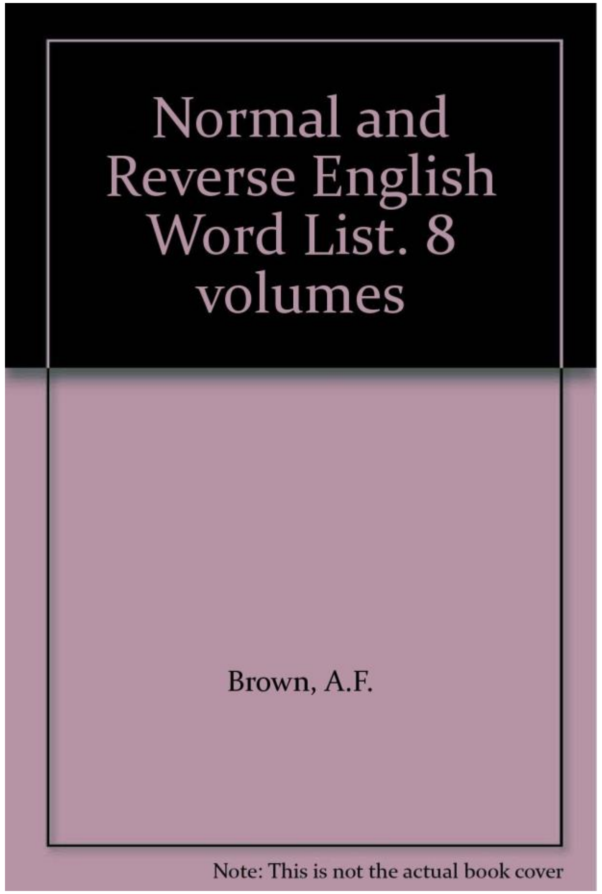
描述: road where cars go very quickly without stopping

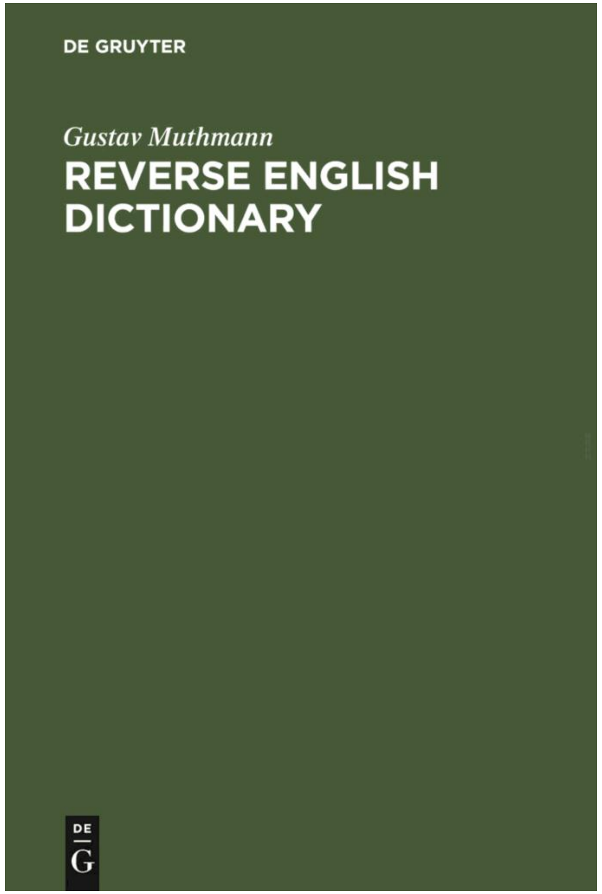
研究背景—反向词典

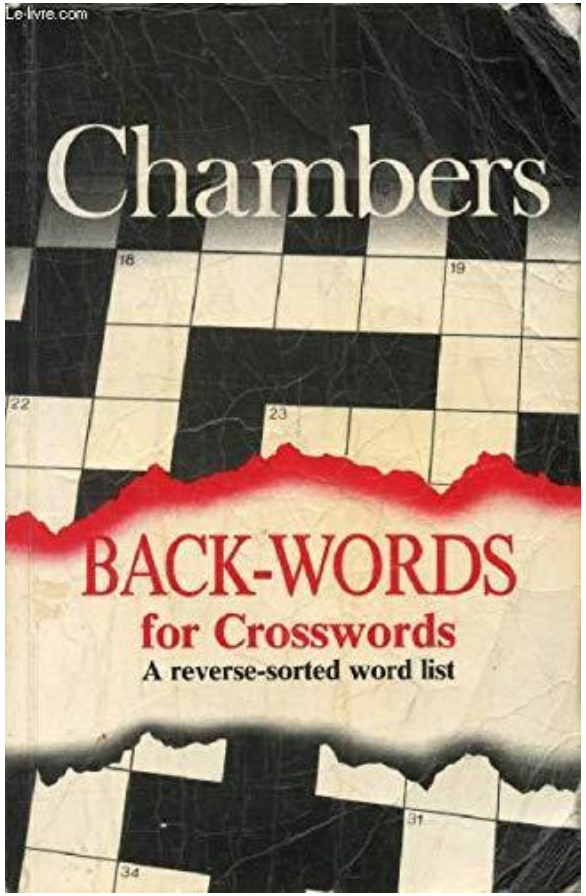
- 反向词典具有重要实用价值
  - 解决舌尖现象（Tip of the tongue，话到嘴边说不出来）
  - 为新语言学习者提供帮助
  - 帮助选词性命名不能（word selection anomia）患者
- 反向词典同样具有自然语言处理研究价值
  - 用于评测句子表示学习质量 [Hill et al. 2016]
  - 有助于包含文本到实体映射的任务，如问答、信息检索 [Kartsaklis et al. 2018]

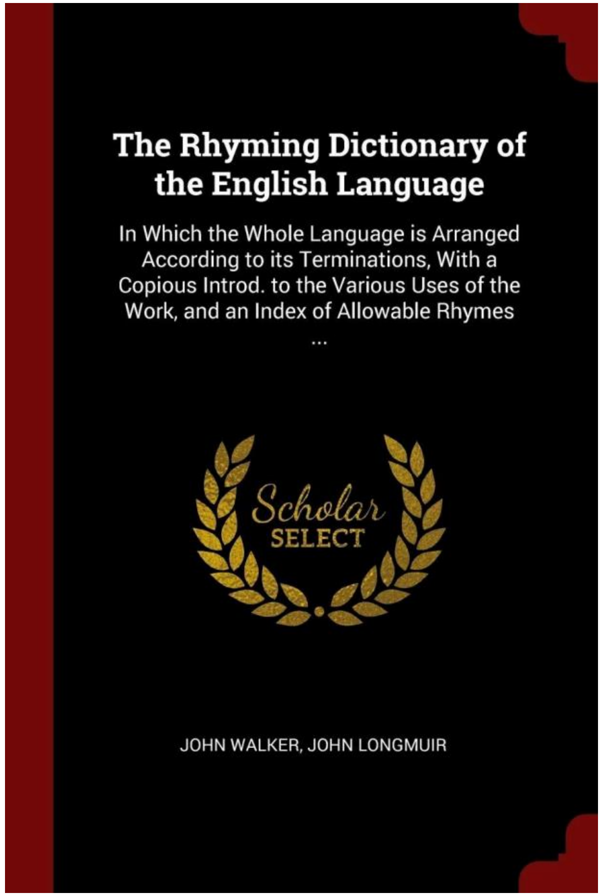
Hill, F.; Cho, K.; Korhonen, A.; and Bengio, Y. 2016. Learning to understand phrases by embedding the dictionary. TACL 4:17 – 30.
Kartsaklis, D.; Pilehvar, M. T.; and Collier, N. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. In Proceedings of EMNLP.

研究背景—现有反向词典


Normal and Reverse English Word List. 8 volumes
Brown, A.F.
Note: This is not the actual book cover


DE GRUYTER
Gustav Muthmann
REVERSE ENGLISH DICTIONARY
DE G


Le-livre.com
Chambers
BACK-WORDS
for Crosswords
A reverse-sorted word list


The Rhyming Dictionary of the English Language
In Which the Whole Language is Arranged According to its Terminations, With a Copious Introd. to the Various Uses of the Work, and an Index of Allowable Rhymes
...
Scholar SELECT
JOHN WALKER, JOHN LONGMUIR

研究背景—现有反向词典

road where cars go very quickly without stopping

OneLook
Thesaurus

Hit Enter or select a word to submit your query.

Show filters | Alphabetize

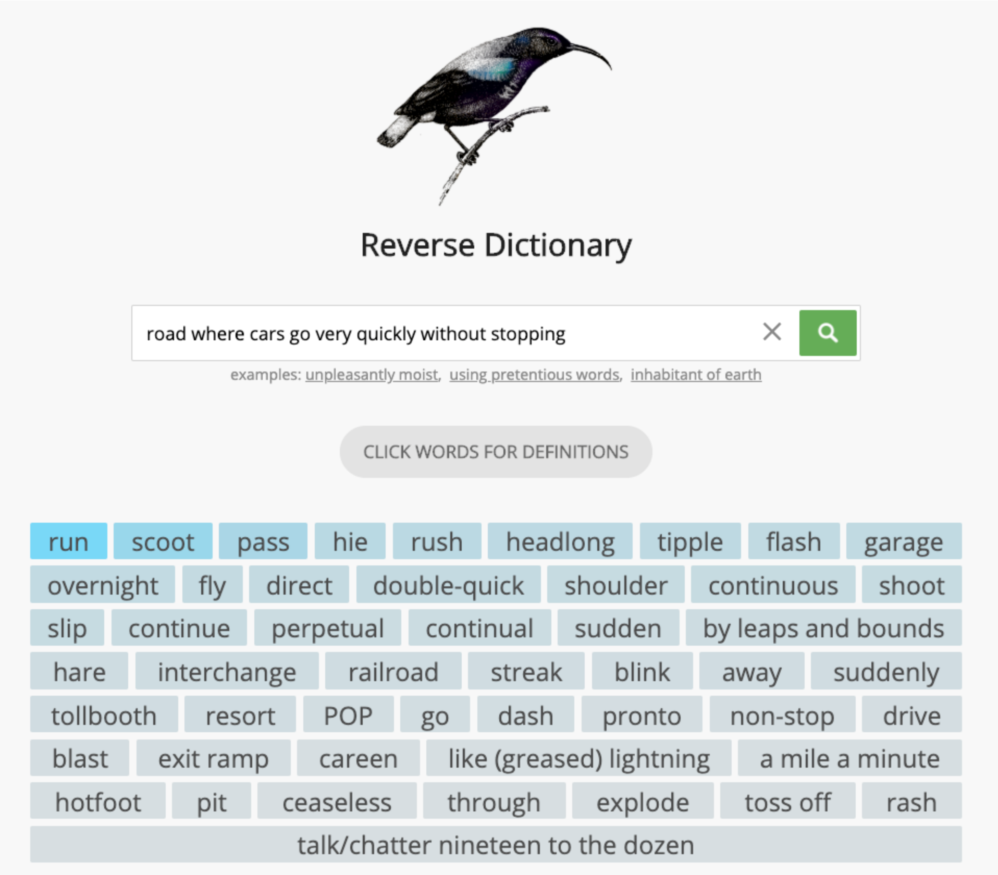
All | Nouns | Adjectives | Verbs | Adverbs

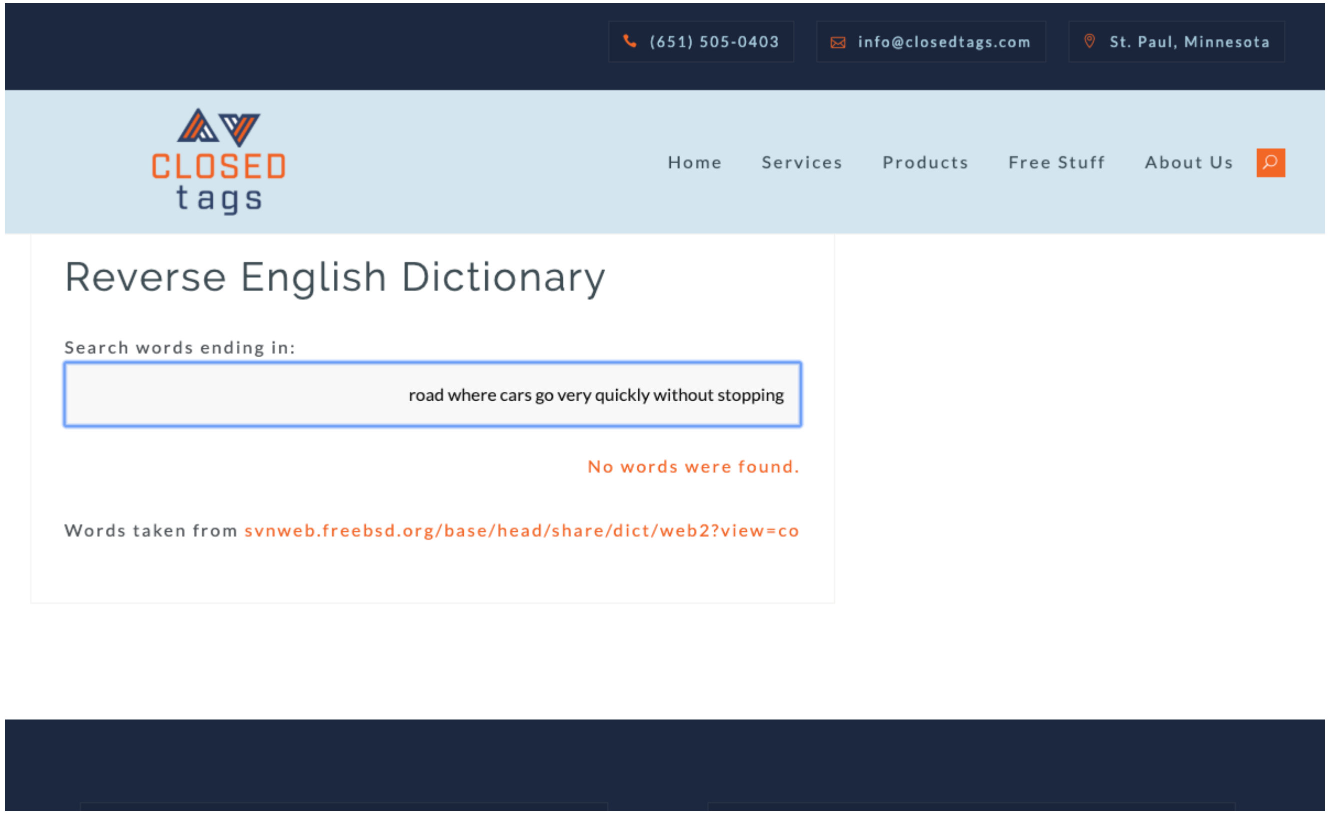
1. station
2. stationing
3. stations
4. pass
5. passed
6. stump
7. stumped
8. stumping
9. stumps
10. shoulder
11. turnout
12. flash
13. run
14. turn
15. turned
16. drag
17. lay
18. stop
19. stops
20. frog
21. part
22. through
23. round
24. skid
25. bridge
26. bridged
27. bridging
28. point
29. passing
30. down
31. take
32. park
33. away
34. layby
35. straight
36. buzz
37. separation
38. fast
39. rope
40. roping
41. runs
42. lay–by
43. jump
44. darts
45. whip
46. pull–off
47. rest area
48. rest stop
49. travel
50. spin
51. walk
52. short
53. scat
54. speed trap
55. leave
56. leaving
57. drop
58. shorts
59. depart
60. wau
61. hydroplane
62. freeway
63. slide
64. hotel
65. pit
66. bomb
67. speed
68. rollercoaster
69. clearway
70. riot
71. railway
72. tracks
73. control
74. catch
75. end
76. course
77. stage
78. gate
79. interchange
80. chicane
81. patt
82. sight
83. drive
84. sights
85. triathlon
86. chamber
87. dozen
88. sledge
89. sledging
90. hell
91. consecutive
92. transfer
93. depot
94. idle
95. stall
96. rounds
97. motorway
98. back
99. plunge
100. buck

Next results >>

https://www.onelook.com/thesaurus/

研究背景—现有反向词典





https://reversedictionary.org/

https://closedtags.com/reverse-english-dictionary/

研究背景—现有反向词典研究

- 基于句子匹配的方法 [Bilac et al. 2004, Zock and Bilac 2004, Méndez et al. 2013, Shaw et al. 2013]
  - 在数据库中存储足够多的词语及其定义
  - 返回与输入描述最相似的词典定义所对应的词语
  - 使用特征工程衡量相似度

反向词典查询输入非常多变，可能与已存储的词典定义有巨大差别！

Bilac, S.; Watanabe, W.; Hashimoto, T.; Tokunaga, T.; and Tanaka, H. 2004. Dictionary search based on the target word description. In Proceedings of NLP.
Zock, M., and Bilac, S. 2004. Word lookup on the basis of associations: from an idea to a roadmap. In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries.
Méndez, O.; Calvo, H.; and Moreno-Armend´ariz, M. A. 2013. A reverse dictionary based on semantic analysis using wordnet. In Proceedings of MICAI.
Shaw, R.; Datta, A.; VanderMeer, D. E.; and Dutta, K. 2013. Building a scalable database-driven reverse dictionary. IEEE Transactions on Knowledge and Data Engineering 25:528540.

研究背景—现有反向词典研究

- 基于神经语言模型的方法[Hill et al. 2016, Morinaga and Yamaguchi 2018, Kartsaklis et al. 2018, Hedderich et al. 2019, Pilehvar 2019]
  - 使用神经语言模型将输入描述编码到词向量空间
  - 返回与输入描述的句子表示最近的词向量对应的词语

存在大量低频词，其词向量效果较差！

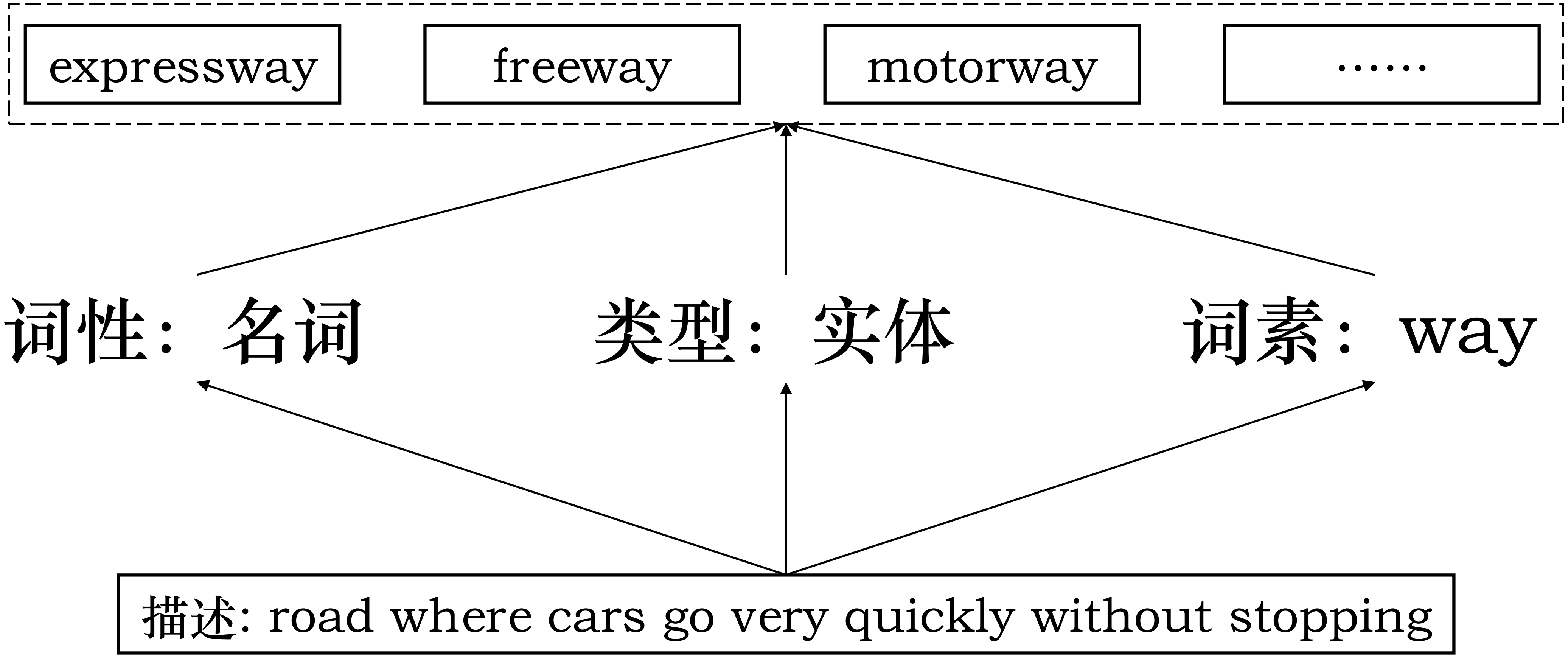
Hill, F.; Cho, K.; Korhonen, A.; and Bengio, Y. 2016. Learning to understand phrases by embedding the dictionary. TACL 4:17 – 30.
Morinaga, Y., and Yamaguchi, K. 2018. Improvement of reverse dictionary by tuning word vectors and category inference. In Proceedings of ICIST.
Kartsaklis, D.; Pilehvar, M. T.; and Collier, N. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. In Proceedings of EMNLP.
Hedderich, M. A.; Yates, A.; Klakow, D.; and de Melo, G. 2019. Using multi-sense vector embeddings for reverse dictionaries. In Proceedings of IWCS.
Pilehvar, M. T. 2019. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. In Proceedings of NAACL.

研究动机—人的推断过程

词性：名词　　类型：实体　　词素：way

描述: road where cars go very quickly without stopping

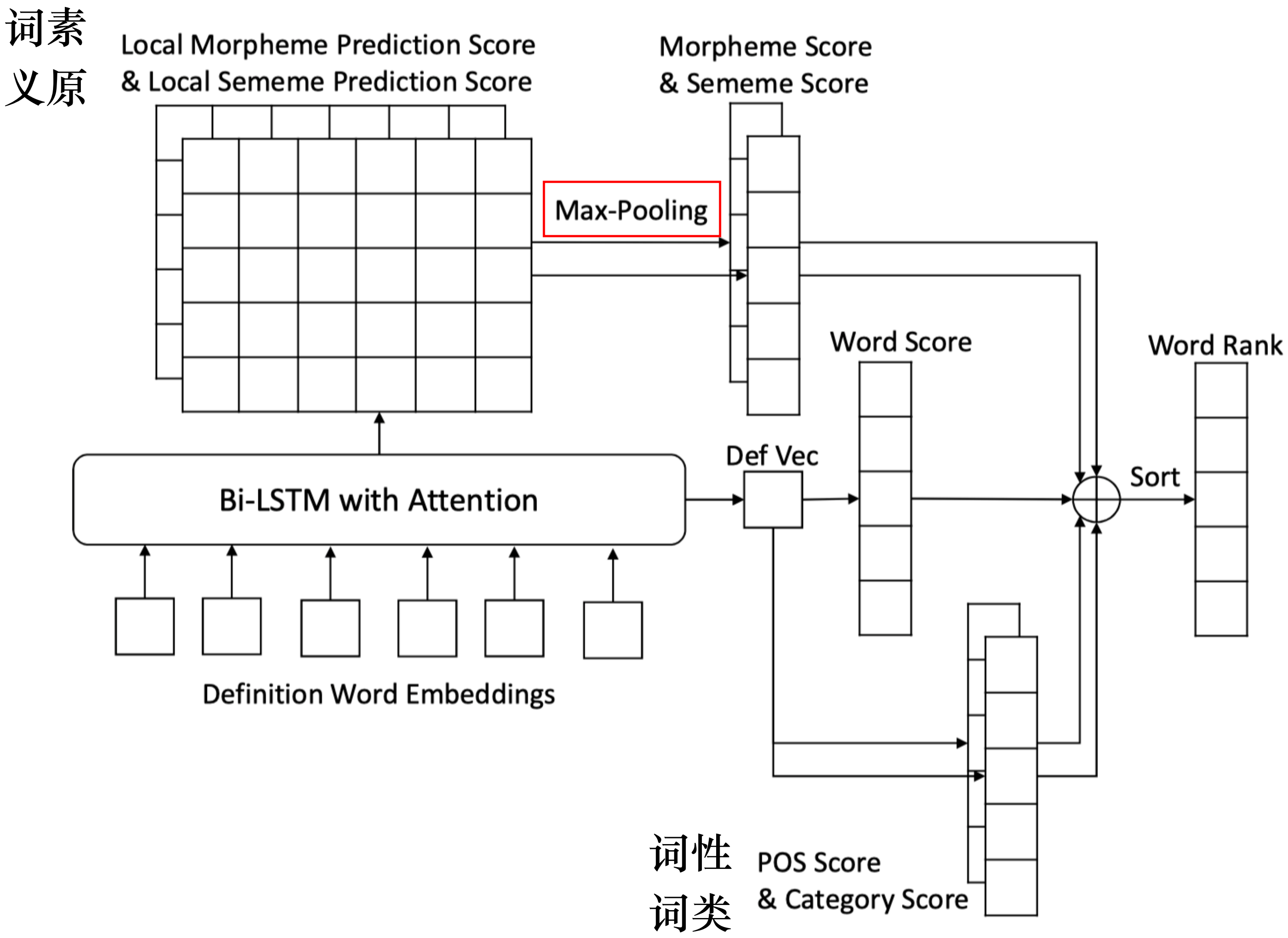
研究动机—人的推断过程


expressway
freeway
motorway
……
词性：名词
类型：实体
词素：way
描述: road where cars go very quickly without stopping

模型

- 多通道反向词典模型——包含多个特征预测器
  - 内部通道——词本身的特征
    - 词性（Part-of-speech）
    - 词素（Morpheme）
  - 外部通道——外部知识库提供的特征
    - 词类（Word Category）——借助WordNet/同义词词林
    - 义原（Sememe）——借助HowNet

模型



词素
义原
Local Morpheme Prediction Score
& Local Sememe Prediction Score
Morpheme Score
& Sememe Score
Max-Pooling
Word Score
Word Rank
Def Vec
Sort
Bi-LSTM with Attention
Definition Word Embeddings
词性
词类
POS Score
& Category Score

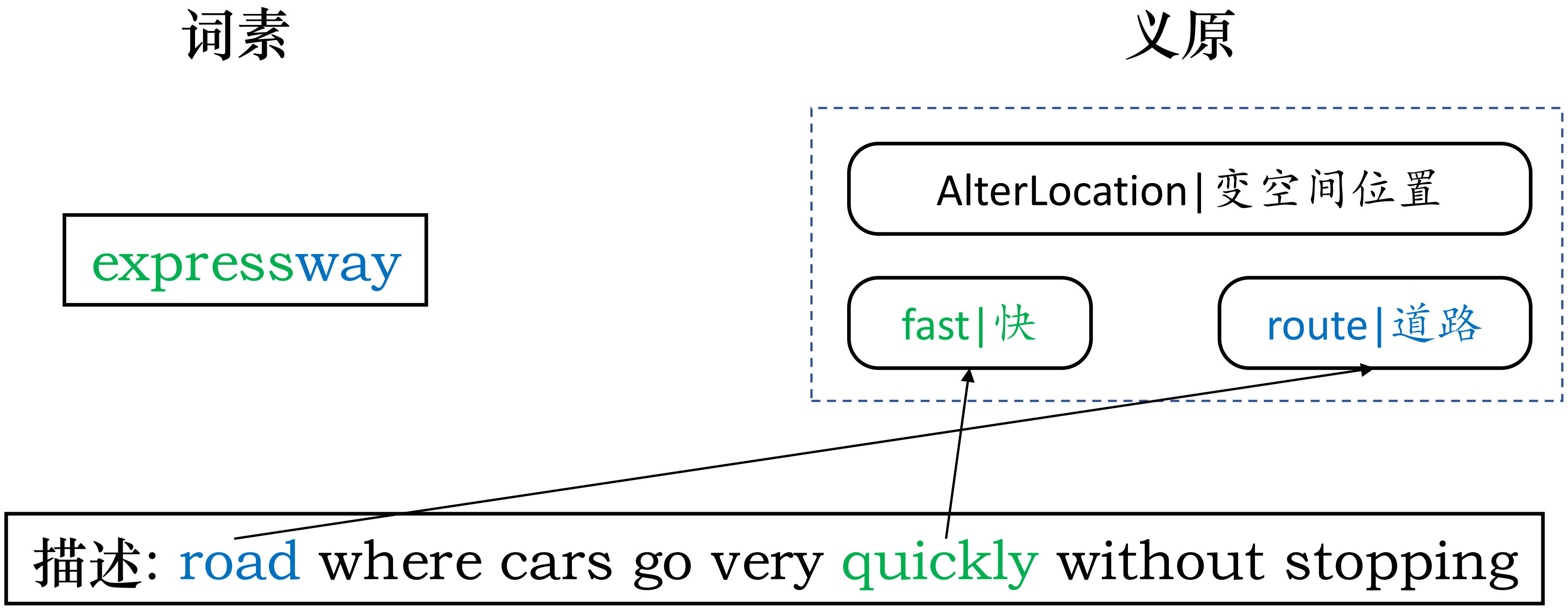
模型

词素

expressway

描述: road where cars go very quickly without stopping

模型


词素
义原
AlterLocation|变空间位置
expressway
fast|快
route|道路
描述: road where cars go very quickly without stopping

实验结果

- 英文数据集

Model	Seen Definition			Unseen Definition			Description		
OneLook	**0**	.66/.94/.95	200	-	-	-	5.5	**.33**/.54/.76	332
BOW	172	.03/.16/.43	414	248	.03/.13/.39	424	22	.13/.41/.69	308
RNN	134	.03/.16/.44	375	171	.03/.15/.42	404	17	.14/.40/.73	274
RDWECI	121	.06/.20/.44	420	170	.05/.19/.43	420	16	.14/.41/.74	306
SuperSense	378	.03/.15/.36	462	465	.02/.11/.31	454	115	.03/.15/.47	396
SOTA MS-LSTM	**0**	**.92/.98/.99**	**65**	276	.03/.14/.37	426	1000	.01/.04/.18	404
BiLSTM	25	.18/.39/.63	363	101	.07/.24/.49	401	5	.25/.60/.83	214
+Mor	24	.19/.41/.63	345	80	.08/.26/.52	399	4	.26/.62/.85	**198**
+Cat	19	.19/.42/.68	309	68	.08/.28/.54	362	4	.30/.62/.85	206
+Sem	19	.19/.43/.66	349	80	.08/.26/.53	393	4	.30/**.64**/.87	218
Multi-channel	16	.20/.44/.71	310	**54**	**.09/.29/.58**	**358**	**2**	.32/**.64/.88**	203

median rank *accuracy@1/10/100* *rank variance*

实验结果

- 中文数据集

Model	**Seen** Definition			**Useen** Definition			**Description**			**Question**		
BOW	59	.08/.28/.56	403	65	.08/.28/.53	411	40	.07/.30/.60	357	42	.10/.28/.63	362
RNN	69	.05/.23/.55	379	103	.05/.21/.49	405	79	.04/.26/.53	361	56	.07/.27/.60	346
RDWECI	56	.09/.31/.56	423	83	.08/.28/.52	436	32	.09/.32/.59	376	45	.12/.32/.61	384
BiLSTM	4	.28/.58/.78	302	14	.15/.45/.71	343	13	.14/.44/.78	**233**	4	.30/.61/.82	243
+POS	4	.28/.58/.78	309	14	.16/.45/.71	346	13	.14/.44/.79	255	5	.25/.59/.79	271
+Mor	1	.43/.73/.87	260	11	**.19**/.47/.73	332	8	.22/.52/**.83**	251	1	.42/.73/.86	227
+Cat	4	.29/.58/.78	319	16	.14/.43/.70	356	13	.16/.45/.77	289	3	.33/.62/.82	246
+Sem	4	.29/.60/.80	298	14	.16/.45/.72	340	12	.15/.45/.75	244	4	.34/.61/.83	231
Multi-channel	**1**	**.49/.78/.90**	**220**	**10**	.18/**.49/.76**	**310**	**5**	**.24/.56**/.82	260	**0**	**.50/.73/.90**	**223**

median rank *accuracy@1/10/100* *rank variance*

实验结果

- 给定先验知识
  - 英文

Prior Knowlege	**Seen** Definition			**Unseen** Definition			**Description**		
None	16	.20/.44/.71	310	54	.09/.29/.58	358	2.5	.32/.64/.88	203
POS Tag	13	.21/.45/.72	290	45	.10/.31/.60	348	3	.35/.65/.91	174
Initial Letter	1	.39/.73/.90	270	4	.26/.63/.85	348	0	.62/.90/.97	160
Word Length	1	.40/.71/.90	269	6	.25/.56/.84	346	0	.55/.85/.95	163

median rank   *accuracy@1/10/100*   *rank variance*

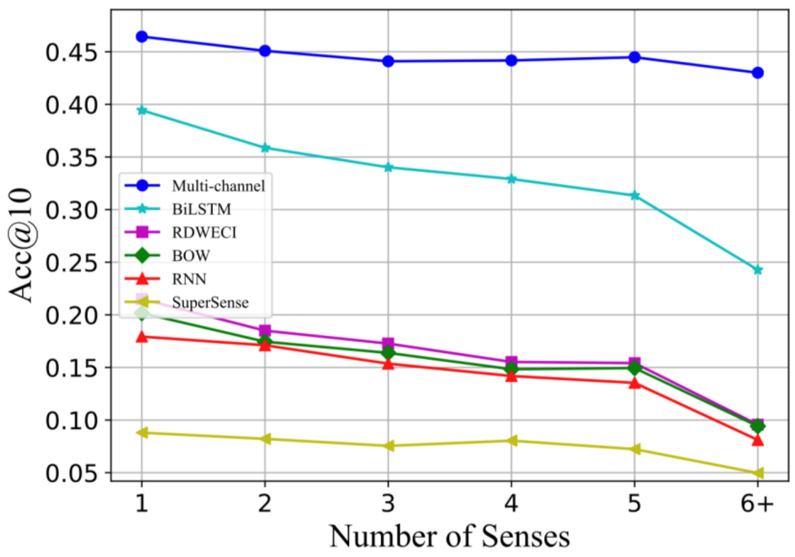
  - 中文

Prior Knowledge	**Seen** Definition			**Useen** Definition			**Description**			**Question**		
None	1	.49/.78/.90	220	10	.18/.49/.76	310	5	.24/.56/.82	260	0	.50/.73/.90	223
POS Tag	1	.50/.79/.90	222	9	.18/.51/.77	307	4	.24/.61/.85	252	0	.50/.74/.90	223
Initial Char	0	.74/.89/.92	220	0	.55/.82/.86	304	0	.61/.88/.93	239	0	.84/.95/.95	213
Word Length	0	.54/.82/.91	217	6	.23/.57/.81	297	3	.32/.68/88	242	0	.62/.85/.94	212

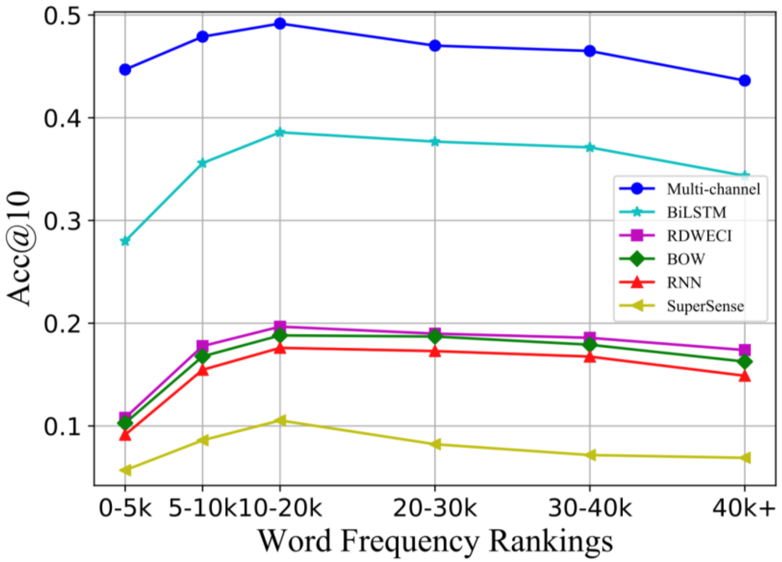
median rank   *accuracy@1/10/100*   *rank variance*

实验结果

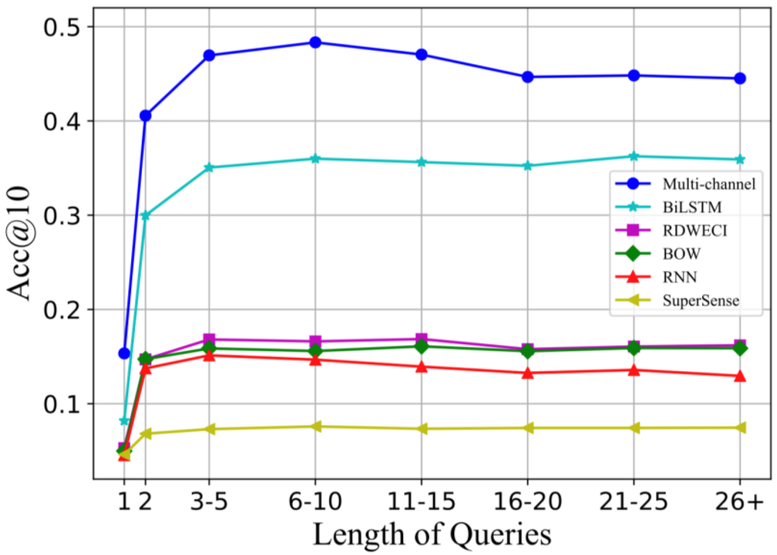
- 模型鲁棒性



义项数



词频



查询长度

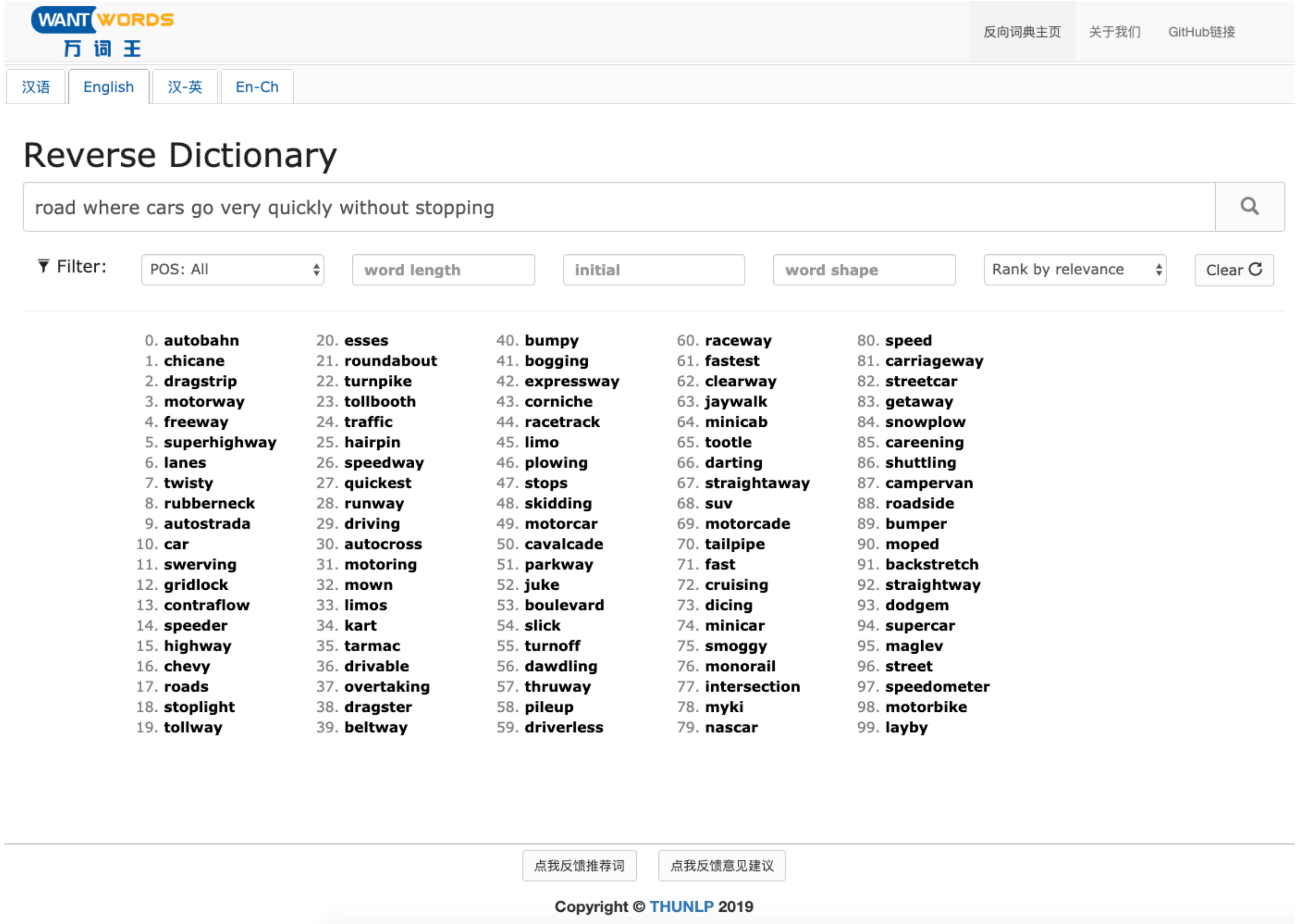
结论

- 主要贡献
  - 基于人的描述-词的推断过程，提出多通道反向词典模型，包含内部通道：词性 、 词素，和外部通道：词类、义原
  - 在真实场景数据集实现了当前最佳性能（state-of-the-art）
- 未来工作
  - 和基于文本匹配的方法相结合，更好的解决极端情况（如输入只有单个词）
  - 迁移到其他任务

论文地址

项目地址

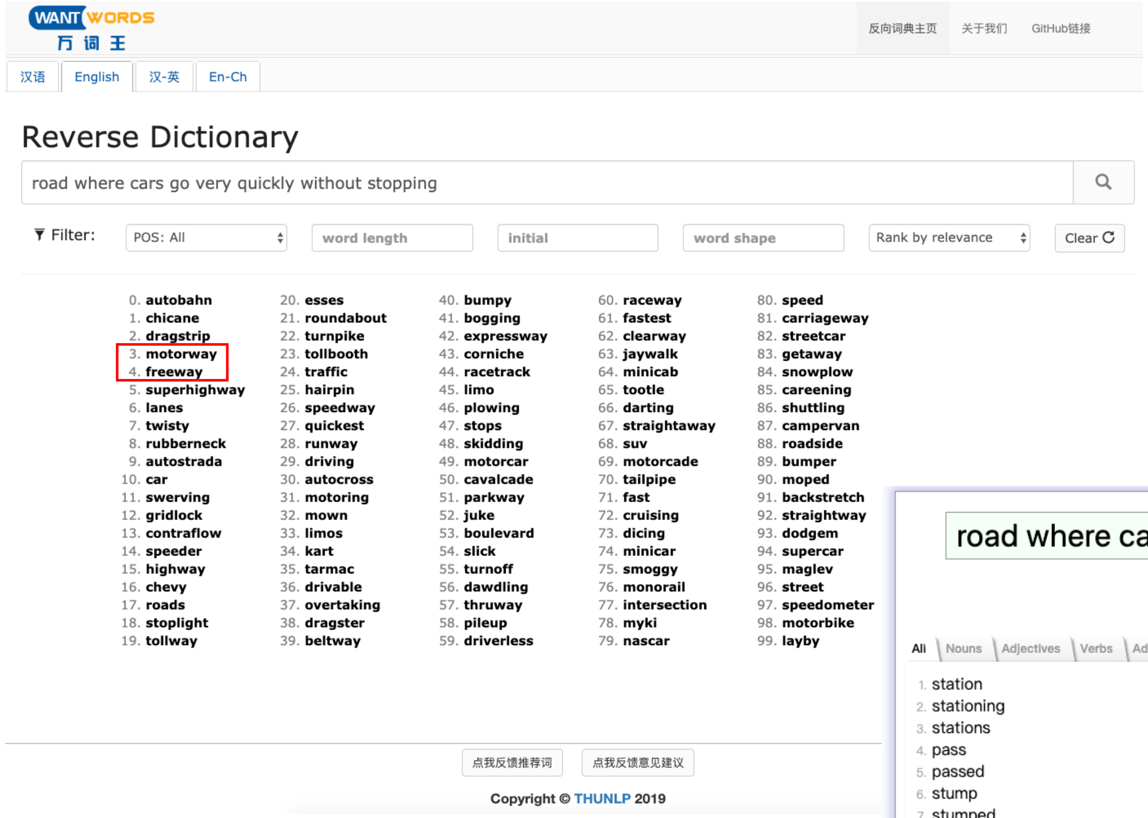
One More Thing

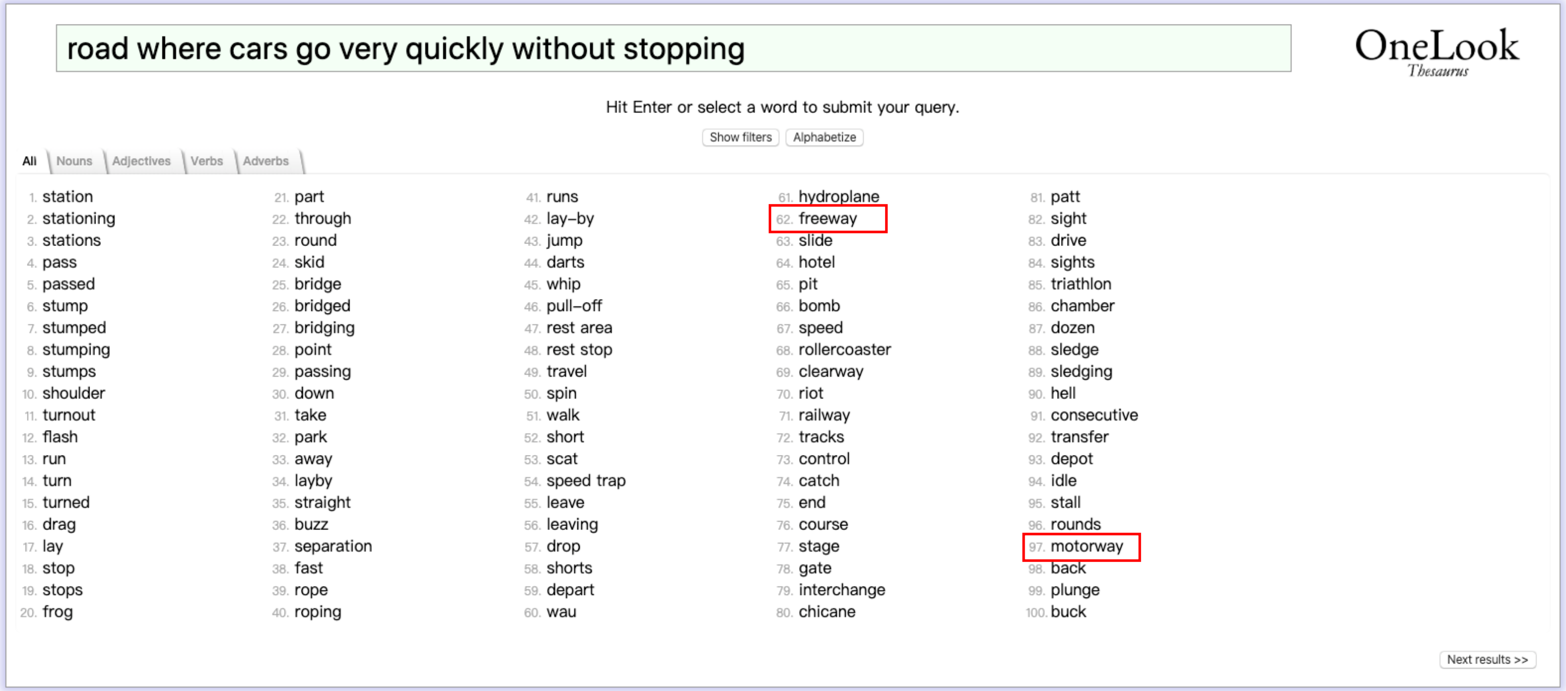


https://wantwords.thunlp.org/

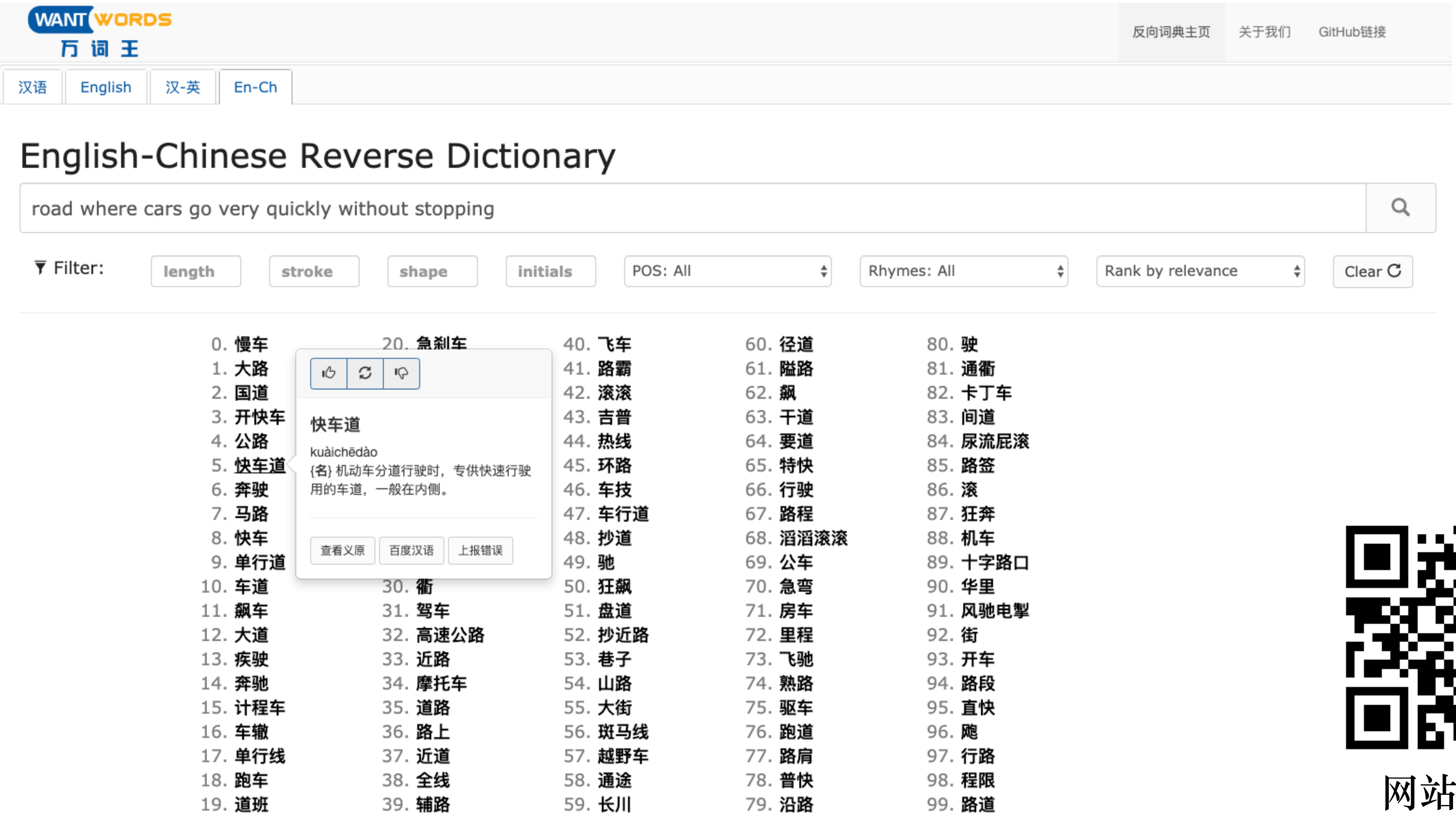
网站地址

One More Thing



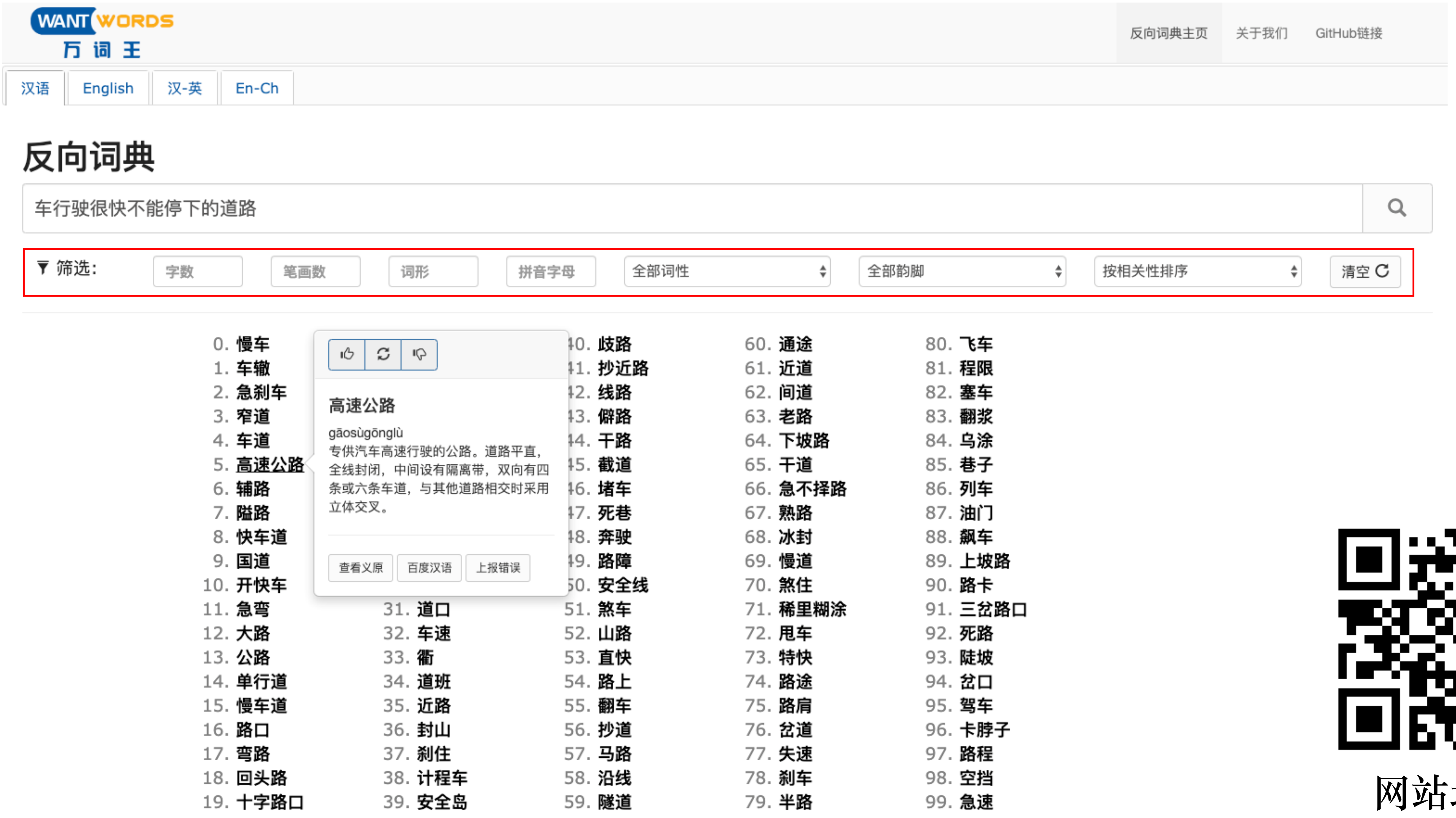


One More Thing



网站地址

One More Thing



网站地址

感谢聆听，欢迎指教！

论文地址

项目地址

网站地址