哈爾濱工業大學
社会计算与信息检索研究中心

基于字信息学习词汇分布的实体上位关系识别

CCKS2016-全国知识图谱与语义计算大会

刘燊，姜天文，秦兵，刘挺

目录

1 引言

1.引言

- ✓ 传统领域命名实体
- ✓ 开放域命名实体
- ✓ 上位词
- ✓ 词汇分布表示

1.引言

- ✓ 传统领域命名实体
  - 主要分为三种：人名、地名、机构名
  - 应用于自然语言处理，无法满足实际需求

1.引言

- ✓ 开放域命名实体
  - 类型更多、更细，具有层次化
  - 难通过人工定义类别体系
  - 使用实体的上位词作为实体的类别

1.引言

✓ 上位词

- 一个语言学概念，它指语义范畴相对较广的词语
  - 如“美洲豹” 是一种“动物” ，则“动物” 就被称为“美洲豹”的上位词
- 抽取上下位关系
  - 基于模式匹配的方法抽取上下位关系，但人工构建的模式仅能处理小部分语言现象，且费时费力
  - 同时Snow等人自动抽取模式的方法对句法分析和语料质量的要求很高，不容易应用到互联网等开放域语料中。
  - **随着深度学习的发展，大量研究基于词汇分布表示开始进行**

1.引言

- ✓ 词汇分布表示
  - 将词语表示成稠密且低维的实数向量，从而使得词语之间可以进行数学运算
  - 可以保留语言的规律性，用于计算词语之间的关系
  - 基于词信息学习词向量表示
  - 未登录词
  - 基于字信息学习词向量表示

哈爾濱工業大學
社会计算与信息检索研究中心

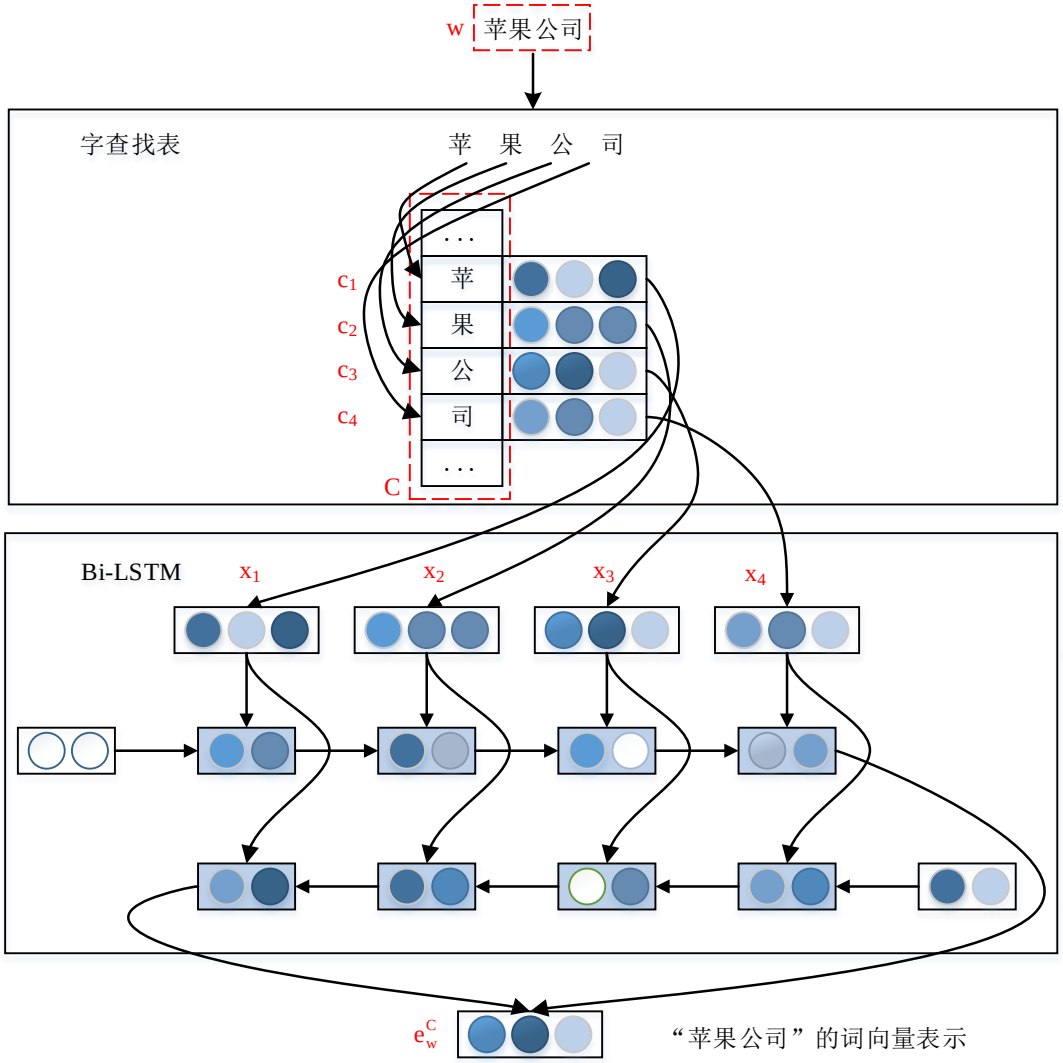
2

基于字信息的词向量学习模型

2.基于字信息的词向量学习模型

- ✓ C2W（character to word）模型
  - Wang等人提出
  - 基于双向LSTM学习词向量
  - 通过学习字之间的信息来组合成词向量的表示

2.基于字信息的词向量学习模型



2.基于字信息的词向量学习模型

基于字信息学习词向量

使用C2W模型来学习字信息

基于字信息重组词向量

2.基于字信息的词向量学习模型

- ✓ 上位关系向量表示
- ✓ 上位关系识别

2.基于字信息的词向量学习模型

- ✓ 上位关系向量表示
  - v(king)−v(queen)≈v(man)−v(women)
  - 两个向量之间的向量差值可以表达出词对之间一定的语义信息
  - 上位关系之间是否也具有类似的现象?

2.基于字信息的词向量学习模型

✓ 上位关系向量表示

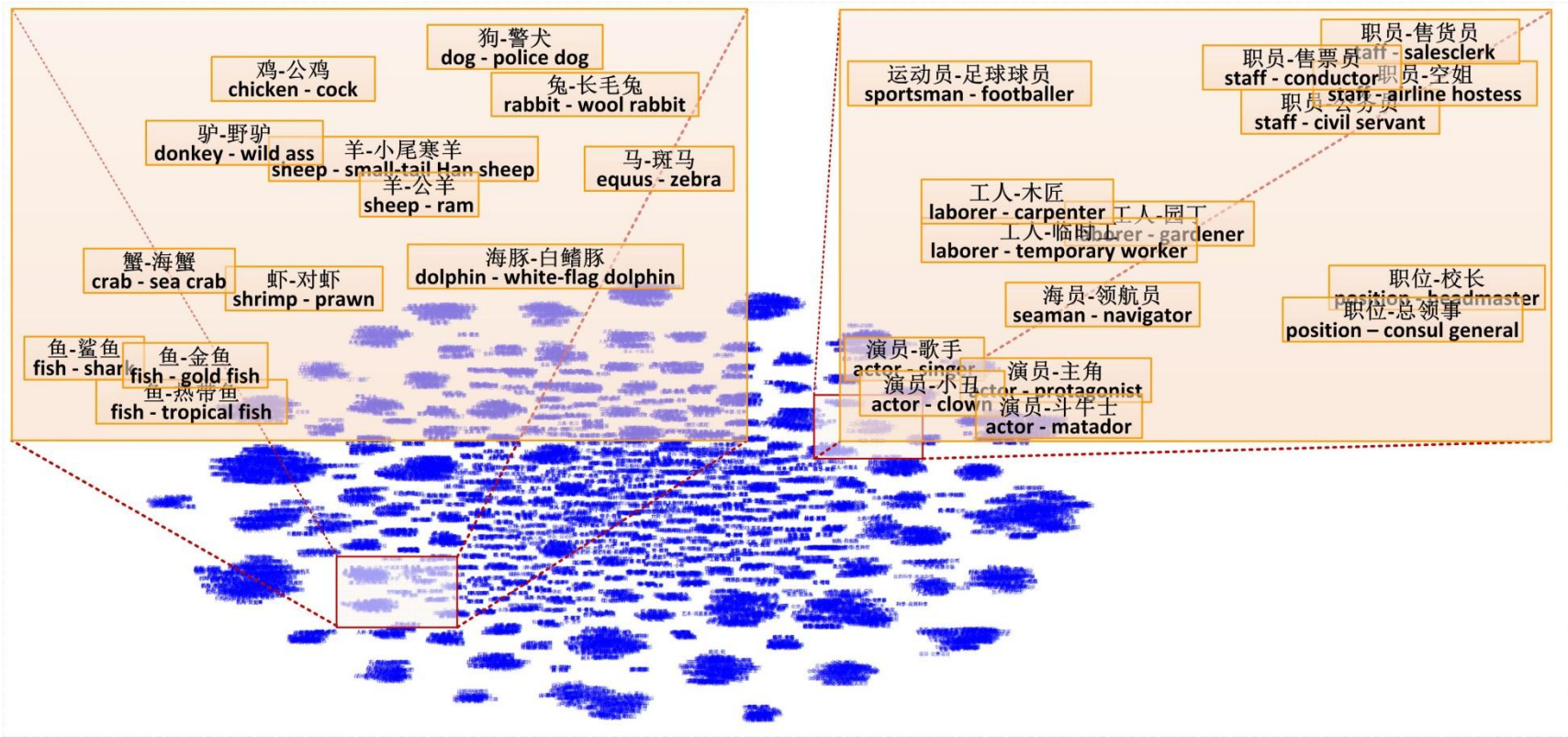
- 上位关系之间是否也具有类似的现象?

序号	实例
1	v(虾)−v(对虾)≈v(鱼)−v(金鱼)
2	v(工人)−v(木匠)≈v(演员)−v(小丑)
3	v(工人)−v(木匠)≈v(鱼)−v(金鱼)

2.基于字信息的词向量学习模型

- ✓ 上位关系向量表示
  - 上位关系之间是否也具有类似的现象?
  - 上下位关系更加复杂，无法简单地使用一个上下位关系向量来表达

2.基于字信息的词向量学习模型



2.基于字信息的词向量学习模型

✓ 上位关系向量表示

- 给定一个词的词向量表示x和它的上位词向量y
- 存在一个矩阵Φ，使得$\mathbf{y} = \Phi\mathbf{x}$
- 最小化均方误差求解下位词到上位词的映射矩阵：
  - $\Phi^* = \arg\min_{\Phi} \frac{1}{N} \sum_{(\mathbf{x},\mathbf{y})} \|\Phi\mathbf{x} - \mathbf{y}\|^2$

2.基于字信息的词向量学习模型

✓ 上位关系向量表示

- 一个具体的下位词往往有多个上位词，因此无法使用单一的映射矩阵来刻画上位关系
- 需要对每一个上位关系向量簇学习一个矩阵映射：

  ➢ $\Phi_k^* = \arg\min_{\Phi_k} \frac{1}{N_k} \sum_{(\mathbf{x},\mathbf{y}) \in C_k} \|\Phi_k \mathbf{x} - \mathbf{y}\|^2$

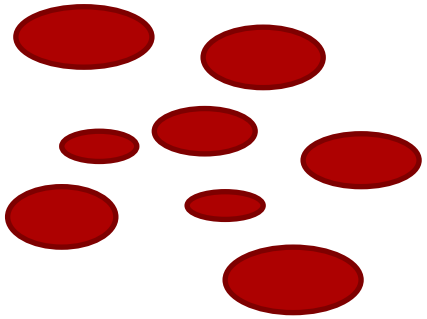
2.基于字信息的词向量学习模型

✓ 上位关系识别

- 上下位关系进行聚类
- 每一个上下位关系簇C_k学习一个向量矩阵Φ_k
- 找出距离y-x向量最近的上下位关系簇Φ_k
- 上位关系显然是存在传递性的

2.基于字信息的词向量学习模型

✓ 上位关系识别



上下位关系进行聚类

每一个上下位关系簇C_k学习一个向量矩阵Φ_k

找出距离y-x向量最近的上下位关系簇Φ_k

3 实验结果与结论分析

3.实验结果与结论分析

- ✓ 词汇分布训练
  - 百度百科中文语料：100多万百科词条，共约3000万句，文件大小4GB左右
  - 分别使用word2vec和C2W模型获得词向量，词向量维度设置为300

3.实验结果与结论分析

- ✓ 使用C2W模型训练所得的词向量，其中部分词的词向量最近5个词结果如下表所示：

词语	相似度	词语	相似度	词语	相似度
中国	1.0000	北京	1.0000	清华大学出版社	-
德国	0.8379	南京	0.9569	出版社	0.7924
美国	0.8144	东京	0.9371	高等学校	0.7742
泰国	0.8134	南北	0.7959	清华大学	0.7664
大国	0.7935	东北	0.7832	师范学院	0.7626
爱国	0.7886	南海	0.7830	理工大学	0.7564

3.实验结果与结论分析

- ✓ 上下位关系簇聚类使用《同义词词林》抽取所得的上下位关系词对数据进行：

关系类型	训练集	开发集	总计
上位-下位关系对词对数	13,718	1,524	15,242

3.实验结果与结论分析

✓ 上位关系识别的两个测试数据集：

关系类型	《同义词词林》数据集	《大词林》数据集
上位-下位关系词对数	2,158	752
无关系词对数	3,250	1,864
总计词对数	5,408	2,590

3.实验结果与结论分析

- ✓ 使用word2vec在《大词林》数据集进行上位关系识别实验结果：

数据	词向量处理方式	未登录词比例	P	R	F1
实体与类别词	无	77.39%	1.0000	0.1607	0.2769
	Avg	33.51%	0.8909	0.3952	0.5475
	Min		0.9787	0.3710	0.5380
	Max		0.9778	0.3548	0.5207
类别词之间	无	15.83%	0.9683	0.3836	0.5496
	Avg	11.15%	0.8289	0.3851	0.5250
	Min		0.9688	0.3780	0.5439
	Max		0.9683	0.3720	0.5374

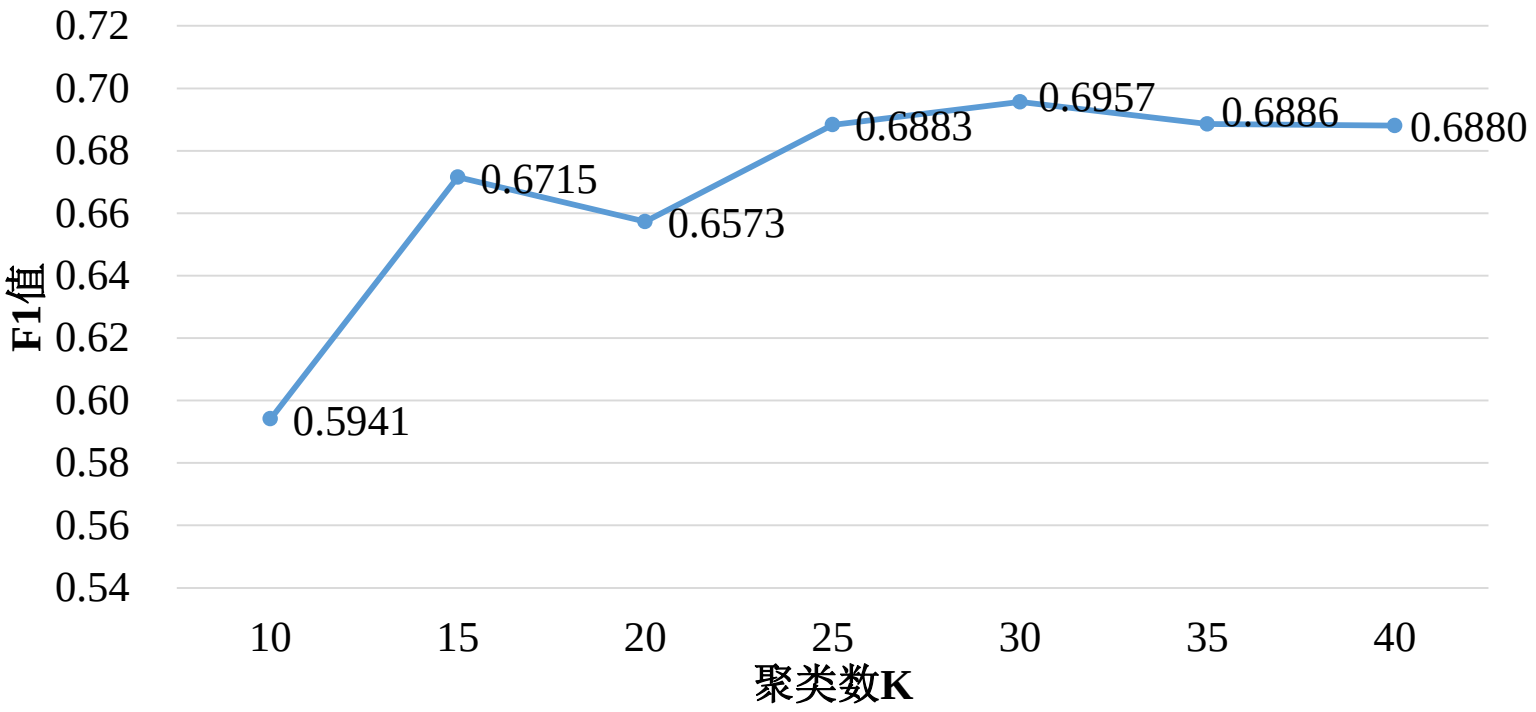
3.实验结果与结论分析

- ✓ 使用word2vec在《大词林》数据集进行上位关系识别结果分析：

  - 未登录词所占比例较大，特别是开放域命名实体与类别词上下位关系部分；
  - 对于原始词语进行分词处理后也还是存在一定量的未登录词；
  - 对于原始词语进行分词处理前后的上位关系识别准确率都较高，基本大于80%，对于部分结果甚至高于95%；
  - 上位关系识别的召回率普遍较低。

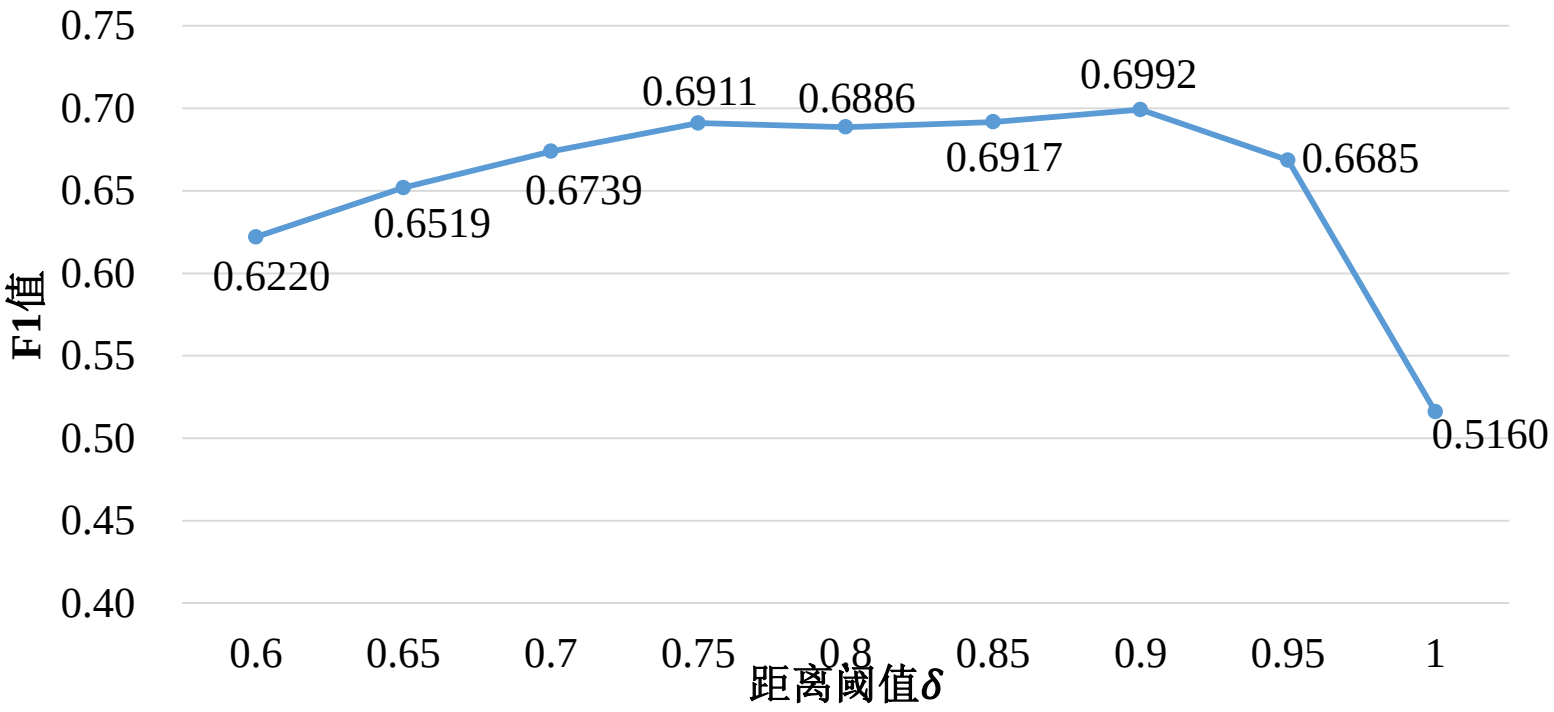
3.实验结果与结论分析

- ✓ 使用C2W模型学习所得词向量作为获得上位关系向量的来源，聚类数目K对结果产生的影响：



3.实验结果与结论分析

✓聚类数目为31时，对距离阈值δ进行了调整：



3.实验结果与结论分析

✓ C2W VS word2vec：

测试数据集	词向量来源	方法	P	R	F1
《同义词词林》数据集	word2vec	M_{Emb}	0.8054	0.6799	0.7374
		$M_{Emb+CilinE}$	0.8059	0.7242	0.7629
		$M_{Emb+CilinE+Wiki}$	0.7978	0.8081	0.8029
	C2W	M_{Emb}	0.7882	0.6282	0.6992
		$M_{Emb+CilinE}$	0.8015	0.6891	0.7411
		$M_{Emb+CilinE+Wiki}$	0.7839	0.7565	0.7700
《大词林》数据集	word2vec	M_{Emb}	0.7609	0.2369	0.3613
		$M_{Emb+CilinE}$	0.7500	0.4772	0.5832
		$M_{Emb+CilinE+Wiki}$	0.7717	0.4805	0.5923
	C2W	M_{Emb}	0.9449	0.3191	0.4771
		$M_{Emb+CilinE}$	0.7927	0.5798	0.6697
		$M_{Emb+CilinE+Wiki}$	0.7935	0.5824	0.6718

4 结束语

4.结束语

- 针对词向量应用中的未登录词问题，本文使用C2W基于字信息的词向量学习模型。

- C2W模型在《同义词词林》所得数据中，上位关系识别结果与word2vec所得效果相当，略低于word2vec。

4.结束语

- C2W模型在《大词林》所得数据中，上位关系识别结果优于使用word2vec所得结果，很大程度上缓解了未登录词的词向量学习问题。

- 未来可以将word2vec与C2W相结合，既缓解未登录词的问题，在词向量的学习上也能够更好地学习词语的语义信息。

参考文献

- 付瑞吉. 开放域命名实体识别及其层次化类别获取[D]. 哈尔滨工业大学, 2014.
- Suchanek F M, Kasneci G, Weikum G. Yago: A large ontology from wikipedia and wordnet[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217.
- Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
- Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]

参考文献

- Snow R, Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery[J]. Advances in Neural Information Processing Systems 17, 2004.
- Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]//HLT-NAACL. 2013: 746-751.
- Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]. In Proceedings of Workshop at ICLR, 2013.
- Fu R, Guo J, Qin B, et al. Learning Semantic Hierarchies via Word Embeddings[C]//ACL (1). 2014: 1199-1209.

参考文献

- Ling W, Luís T, Marujo L, et al. Finding function in form: Compositional character models for open vocabulary word representation[C]. EMNLP, 2015.
- Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.
- Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 13-16.

哈爾濱工業大學
社会计算与信息检索研究中心

谢谢各位聆听!