**基于词表示的查询日志中实体链接方法研究**

摘 要

搜索引擎最首要的任务是理解用户的意图，即理解查询日志，其中的核心步骤是识别查询日志中的实体并将其链接到目标知识库，然而当前搜索引擎的表现仍有提升的空间，实体链接的方法需要进一步改进。同时，当前各种词表示技术被应用到自然语言处理的各项任务中，并取得了不错的结果。因此本文尝试将词表示的技术应用于查询日志中的实体链接任务。

本文首先对词表示模型进行探究与改进，在已有的词表示模型的基础上提出了忽略虚词的词表示模型和语义增强的词表示模型，并在词语类推和词义消歧任务中分别对其进行测试，实验结果表明：忽略虚词的词表示模型，可以在语义相关的任务上有更高的准确率，而在语法相关的任务上准确率有所降低；将《知网》中的义原知识融入词表示的训练模型中会提升模型的语义表示能力，相比较其他无监督的方法，语义增强的词表示模型在中文词义消歧任务中获得了更高的准确率。然后是实体链接方法的研究，并将不同的词表示模型应用到查询日志中实体链接的任务中。对于候选实体的生成模块，本文采用基于搜索引擎和实体词典的方法；对于实体消歧模块，本文利用候选实体和实体指称之间的相似度进行消歧，其中包含基于词袋模型的相似度计算方法和改进的基于词袋模型的相似度计算方法。

本文工作的创新点在于：探究虚词在词表示模型中的影响，通过实验发现，忽略虚词后，在语法和语义任务上对模型有不同的影响；首次将《知网》中的语义知识融入到中文词表示的模型训练中，增强了模型的语义表示能力；在词袋模型的基础上，我们引入候选实体本身和实体指称本身之间的相似度，并应用于实体消歧模块，同时将改进的词表示模型应用于实体链接工作中。

通过实验发现：相比于基于词袋模型的相似度计算方法，加入候选实体本身和实体指称本身之间的相似度后，实体链接的平均F值有较大的提升，从0.365提升到0.592。同时，将忽略虚词的词表示模型和语义增强的词表示模型应用到实体链接任务后，相比较普通的词表示模型，实体链接的平均F值分别由0.592、0.547提升到0.596和0.560。

关键词：实体链接；词表示；词向量；查询日志；《知网》；虚词

STUDY ON ENTITY LINKING METHODS IN SEARCH QUERIES BASED ON WORD REPRESENTATION

Abstract: The key process of understanding a query is to recognize entities and link them to a defined Knowledge Base. While the method of entity linking is still need to be improved to boost the performance of a search engine. Meanwhile, various word representation techniques are applied in natural language processing tasks, and even achieved state of art results. Therefore, this paper try to apply word representation techniques to entity linking in search queries.

This paper proposed a word representation model that without function word and a word representation model that injected with semantic knowledge. Then we tested them on a word analogy task and a word sense disambiguation task separately. The results show that the accuracy of model has an increase on semantic related task and a decrease on syntactic related task after neglecting function word. In addition, the model has a better semantic representation after injecting sememe knowledge of *HowNet*, and achieved a higher accuracy in a Chinese word sense disambiguation task compared with another unsupervised method. Then we did some researches on the method of entity linking, and applied different models to entity linking. For entity generation module, we used a method based on search engine and dictionary. As for entity disambiguation module, we used two methods to compute similarity between candidate entity and entity mention for disambiguation, including method based on bags-of-words and methods based on improved bags-of-words.

Compared with the method based on bags-of-words, the average F-value of entity linking improved significantly after introducing the similarity between candidate entity itself and entity mention itself, which gained from 0.365 to 0.592. Meanwhile, with neglecting function word model and semantic strengthened model, the average F-value also increased from 0.592, 0.547 to 0.596 and 0.560 separately.

Key words: entity linking; word representation; word embedding; search queries; *HowNet*; function word

目 录

[1 绪论 1](#_Toc452744787)

[1.1 研究背景和意义 1](#_Toc452744788)

[1.2 研究现状 2](#_Toc452744789)

[1.3 本文的工作和创新点 4](#_Toc452744790)

[1.4 论文组织结构 5](#_Toc452744791)

[2 原理和知识库 7](#_Toc452744792)

[2.1 词的表示学习 7](#_Toc452744793)

[2.2 知识库 11](#_Toc452744794)

[2.3 语义增强的词的表示学习 13](#_Toc452744795)

[3 搜索查询日志中的实体链接 15](#_Toc452744796)

[3.1 实体链接流程 15](#_Toc452744797)

[3.2 数据处理 16](#_Toc452744798)

[3.3 词表示模型的训练 18](#_Toc452744799)

[3.4 候选实体的生成 23](#_Toc452744800)

[3.5 实体消歧 24](#_Toc452744801)

[4 实验 27](#_Toc452744802)

[4.1 评价准则 27](#_Toc452744803)

[4.2 实验数据 28](#_Toc452744804)

[4.3词表示与实体链接 29](#_Toc452744805)

[4.4 忽略虚词的词表示与实体链接 30](#_Toc452744806)

[4.5 语义增强的词表示与实体链接 33](#_Toc452744807)

[5 总结和展望 37](#_Toc452744808)

[5.1 总结 37](#_Toc452744809)

[5.2 未来工作 37](#_Toc452744810)

[参考文献 39](#_Toc452744811)

[致谢 43](#_Toc452744812)

[已发表文章目录 44](#_Toc452744813)

# 

插 图

[**图 1 CBOW 模型结构图 9**](#_Toc448006673)

[**图 2 Skip-gram模型结构图 10**](#_Toc448006674)

[**图 3 语义增强的词表示模型结构图 13**](#_Toc448006675)

[**图 4 查询日志中实体链接流程图 16**](#_Toc448006676)

[**图 5 义原的分布信息 20**](#_Toc448006677)

[**图 6 单义原词中的义原表示能力 21**](#_Toc448006678)

[**图 7 多义原词中的义原个数统计图 21**](#_Toc448006679)

[**图 8 “材料”的消歧实例 22**](#_Toc448006680)

[**图 9 候选实体的生成图 24**](#_Toc448006681)

[**图 10 忽略虚词后词表示模型在语义子集任务中准确率的变化 32**](#_Toc448006682)

[**图 11 忽略虚词后词表示模型在语法子集任务中准确率的变化 32**](#_Toc448006683)

表 格

[**表 1 主流词表示模型的对比 7**](#_Toc448613248)

[**表 2 词语的共现概率以及比值信息 11**](#_Toc448613249)

[**表 3 《知网》中概念的描述实例 12**](#_Toc448613250)

[**表 4 实体词典的构建实例 17**](#_Toc448613251)

[**表 5 虚词词性列表 19**](#_Toc448613252)

[**表 6 实验数据样例 28**](#_Toc448613253)

[**表 7 词表示模型中训练语料的统计信息 29**](#_Toc448613254)

[**表 8 不同的相似度计算方法以及评测系统的对比实验 29**](#_Toc448613255)

[**表 9 词表示模型在词语类推任务中的对比实验 31**](#_Toc448613256)

[**表 10 不同相似度计算方法的实体链接结果 33**](#_Toc448613257)

[**表 11 词义消歧任务中的对比实验 33**](#_Toc448613258)

[**表 12 词义消歧的实验结果 34**](#_Toc448613259)

[**表 13 “钱”的部分消歧结果与分析 35**](#_Toc448613260)

[**表 14 普通词表示模型与语义增强的词表示模型的对比实验 35**](#_Toc448613261)

# 1 绪论

## 1.1 研究背景和意义

互联网中的数据随着网络技术的发展正在变得愈加庞大，其中大部分的数据都是以自然语言的形式存在。然而自然语言是充满歧义的，尤其像命名实体。命名实体是人名、地名、机构名以及其他以名称为标示的实体。一个命名实体可能有多个名称，而且一个名称可能会指向多个命名实体。因此正确识别和理解命名实体是自然语言理解的基础工作。

### 研究背景

由于每天都有许多新的事实产生，人们需要借助信息抽取的方法对已有的知识库进行扩充，这就需要将实体指称与知识库中的实体进行对应。而实体链接就是将给定的实体指称与到指定知识库中的实体进行匹配和映射。一般可以分为命名实体的识别，候选实体的生成，实体消歧等模块。

同时实体链接与词义消歧的任务相类似。词义消歧是在一段上下文中根据词语的词义目录确认一个词语的词义，但是实体链接是根据知识库来确定一个命名实体的语义，词义消歧任务中词语的词义目录是完整的，而知识库有可能是不完整的，即部分实体并不存在与知识库中。此外，相比词义消歧，实体链接任务中实体指称有更多的可能性，因此难度也更大。

随着网络的发展与普及，人们可以随时随地通过搜索引擎搜索问题，解决疑惑，搜索引擎已经成为人们生活中获取知识和信息的重要途径，极大的方便了人们的生活。虽然大部分情况下搜索结果能够满足人们的需求，但是很多时候仍得不到满意的答案，所以搜索引擎的体验仍然需要进一步地提高。而搜索引擎最首要的任务就是理解用户的意图，即理解查询日志，其中的核心步骤是识别查询日志中的实体并将其链接到目标知识库。

### 1.1.2 研究意义

作为自然语言处理中一项重要的基础性研究，实体链接被广泛应用到上层的句法分析、机器翻译、信息抽取、信息检索、文本分析、问答系统、构建知识图谱等应用中。

当前实体链接的任务多是基于长文档的实体链接，即实体指称出现在一篇文档中，因此有大量的上下文可以利用，为实体链接提供了丰富的特征，因此难度相对较低。与此对应的是针对短文本的实体链接，像新浪微博，Twitter,搜索中的查询日志等领域，一方面，实体指称所在的上下文比较短，可用特征比较少，另一方面，由于属于网络用语，规范性比较差，使得实体链接的难度增加。

搜索引擎是最直接服务于群众的应用之一，提高查询日志语句中的实体链接的准确率可以提升互联网搜索引擎的信息过滤能力，进而改善用户体验，更好地服务于用户。

## 1.2 研究现状

实体链接方法中既包括基于规则的方法，也有基于统计的有监督机器学习和无监督机器学习，并取得了一定的成果。随着计算机硬件水平的提升，计算成本迅速下降，算法运行速度显著提高，同时大数据时代的到来带来了海量的数据，可以训练更准确的模型。这些条件都促进了词表示方法的发展，而且已经被广泛应用到各项自然语言处理的任务中，并在个别任务中获得了最优结果。近年来已有科研工作者意识到词表示技术以及深度神经网络模型的前景，开始将其应用到实体链接任务中，并提升了链接准确率。

### 1.2.1 词表示

近些年，各种基于神经网络语言模型的词表示方法层出不穷，其中最经典的属于Bengio等人第一次提出的神经网络语言模型（NNLM）[1]，在训练语言模型的基础上同时得到了词的表示。Mikolov在神经网络语言模型的基础上进行简化，提出了Skip-gram和CBOW模型[2,3,4]，简单而又高效。另外一种主流的词表示方法是基于矩阵分解的技术，其中Pennington等人最近提出的GloVe模型[5]同Skip-gram和CBOW模型一样，被广泛应用于各种自然语言处理任务中。部分科研工作者在它们的基础上进行优化和改造，同样取得了不错的结果。

尹文鹏等利用各种词表示模型构建了一个混合词表示模型 [6]。选择的是当前最主流的CW, HLBL, Huang, Word2vec, GloVe五个模型，这样做的优点就是可以增强词表示模型的表示能力，增加模型的覆盖面。Yu等借助具有相关语义关系的词语对进行训练[7]，而且该语义关系需要保证一定的质量和数量，通过一个词语的表示来预测另外一个相关词语的表示，称为关系限制模型（Relation Constrained Model, RCM），并将CRM和CBOW模型进行线性加权的结合。Johansson等借助一个语义网络的网络结构信息来提升模型的表示能力[8]， 其中，词义的表示由词条的表示和语义网络中的邻居节点的词表示混合而成，而且其中的邻居节点也是选择最近的、有直接关系的节点，并保持一定的权重。Chen等将词义表示应用于词义消歧中，同时词义消歧的结果再用来优化词义的表示[9],他们利用WordNet中的注释信息对词义表示进行初始化，其中注释中只利用名词、动词、形容词和副词，而且注释中的词和目标词之间的相似度大于一定的阈值才被选择，选定词的词向量的均值作为词义的初始表示值。利用消歧的结果，在Skip-gram模型的基础上修改目标函数，得到词义表示。Ling等认为目标词周围的上下文词语并不是同等重要的，而且主流模型都缺少词序信息，于是在词袋模型的基础上进行扩展，提出了CBOW with Attention模型，对于上下文中的每个词都是有权重的[10]。Faruqui等利用外部知识库中的语义关系，借助图模型对词向量进行了翻新改进[11]。

### 1.2.2 实体链接

传统的实体链接可以分为候选实体的生成和实体消歧两个模块。

候选实体的生成主要有基于词典的方法和基于搜索引擎的方法。基于词典的方法是目前主流的生成候选实体的方法，词典一般由维基百科知识构建而成，最直接的方法是如果词典中存在与实体指称完全匹配的实体，则作为候选实体。Dredze等利用了实体指称与维基百科的标题之间相互匹配的方法和字符相似度算法生成候选实体[12]。Han和Zhao利用谷歌API获取对短文本的搜索结果，保留结果中的维基百科页面，并将其中的实体作为候选实体[13]。Meng等人则利用百度来搜索实体指称+“维基百科”和实体指称+“中文维基百科”，将搜索结果中的维基百科页面里的实体作为实体指称的候选实体[14]。

实体的消歧主要依靠特征信息，Zheng等利用编辑距离来获取实体指称与实体之间字符相似度[15]。Hoffart 等利用词语在维基百科中出现的频率信息定义流行度特征[16]。Zhang等对实体指称的上下文词语构建词袋模型[17]。Shen等用维基百科页面中的概念信息向量来表示文档的语义信息[18]。Zheng 等利用余弦相似度[15]、Hoffart等利用KL距离[16]、Kulkarni等利用Jaccard相似度[19]来计算向量之间的相似度。Shen等融合实体的流行度、语义相关度、语义相似度和全局主题共现信息，利用支持向机模型对候选实体进行排序[18]。Han和Zhao利用基于向量空间模型的无监督方法，计算实体指称与候选实体之间的相似度，进而得到目标链接实体[13]。Shen等增加了同一条Twitter中实体之间的相关性以及用户的兴趣作为特征，提出了一个基于图模型的实体链接系统[20]。Meng等利用新浪微博中微博用户名的属性作为特征，实体链接的准确率由80.07%提升到84.35%[14]。

词表示的方法不需要过多的特征工程，避免了人为地引入错误，而且在没有标注的语料上就能够训练出效果不错的词向量。同时，开源的词向量训练模型简单高效，更使得词表示的方法被广泛应用到各项自然语言处理任务中，并且了取得不错的效果。部分学者开始尝试将其应用到实体链接的任务中。

Wang等利用知识图谱和语料库同时训练出词语和实体的词向量，在训练过程中，利用实体名和维基百科的锚点来对齐实体向量和普通词向量，保证其在同一个向量空间[21]。

He等利用深度神经网络开展实体消歧工作，在预训练阶段，首次提出了利用堆栈式降噪自动编码器来表示初始化文档表示，然后在给定相似度度量方法的条件下，优化文档和实体的表示[22]。

Ling等在词袋模型的基础上引入依存特征，将相关性比较低的词语从词袋中去除，减少了词袋中不相关的词语给消歧带了噪音。同时利用各种知识库构建一个低维的向量空间来解决数据稀疏问题[23]。

## 1.3 本文的工作和创新点

本文的研究重点可以分为两部分，一部分是对词表示模型的研究与改进，另一部分是将词表示模型应用到搜索查询日志中的实体链接任务。对于词表示模型的研究与改进，一方面是研究虚词对词表示模型的影响，另一方面是利用外部语义知识增强词语的语义表示能力。针对如何将词表示的方法应用到实体链接任务中，首先选择最直接的基于词袋模型的方法，在此基础上引入候选实体本身与实体指称本身之间的相似度提升实体链接的准确率。

本文中实体链接的主要流程包括：

1. 数据准备：获取用于训练词表示模型的训练语料，以及其他相关语料和数据集，进行分词，词性标注操作，并训练词表示模型；从《知网》中抽取和统计义原信息。
2. 候选实体的生成：采用基于搜索引擎和词典的方法生成候选实体。
3. 普通词表示模型与实体消歧：利用词表示模型，首先选择基于词袋模型的相似度计算方法，即候选实体的摘要信息和候选实体所在的查询语句之间的相似度进行实体消歧，然后是基于改进的词袋模型的相似度计算方法，即引入候选实体本身与实体指称本身之间的相似度进行实体消歧。
4. 忽略虚词的词表示模型与实体消歧：在词表示模型的训练过程中，删除训练语料中的虚词，然后在词语类推任务中对模型进行评估，并将其应用到实体消歧中。
5. 语义增强的词表示模型与实体链接：将《知网》中的具有语义表示能力的义原信息融入到词表示的训练模型中，在类似的词义消歧任务中进行验证并将其应用到实体消歧中。

本文工作的创新点包括：

1. 探究虚词在词表示的模型中的影响，通过对词表示模型的评测发现，不论是CBOW模型还是Skip-gram模型，去掉虚词后的词表示模型具有更好的语义表示能力，在语义子任务中准确率得到提升，相反，在语法子任务中的准确率普遍降低。同时，对于GloVe模型，忽略虚词对语义和语法任务的影响都很微小。因此在不同的自然语言处理任务中可以针对性的选择是否需要忽略虚词。
2. 本文首次将《知网》中的义原知识融入到中文词表示的模型训练中，并将其应用到词义消歧和实体消歧两项任务中，实验结果表明该方法简单有效，而且能够提高实体链接的准确率。
3. 在词袋模型的基础上，即利用实体指称所在的查询语句和候选实体的摘要之间的相似度，我们引入候选实体本身和实体指称本身之间的相似度，共同应用于实体消歧模块。

## 1.4 论文组织结构

本文共分为五个章节：

第一章为绪论部分，介绍实体链接工作的研究背景和意义、相关研究工作进展，简述本文的研究方法、思路和研究内容，并列举出本文采用方法的创新之处。

第二章主要介绍本文所采用方法的原理和使用到的知识库。首先是对词表示模型的介绍，其中重点描述CBOW、Skip-gram和GloVe三个模型的原理。然后介绍《知网》的内容与知识结构，应用到词表示模型的“义原”概念以及维基百科。最后介绍语义增强的词表示模型，如何将义原信息融入到普通的词表示模型中增强词语的语义表示能力。

第三章主要描述本文采用的实体链接方法。介绍实体链接的整个流程，并详细介绍每个环节的具体实现，包括：数据处理、词表示模型的训练与改进、候选实体的生成和实体消歧。

第四章是实验部分，介绍了评价方法、实验所用的数据以及三个对比实验：基于不同相似度计算方法的实体链接；忽略虚词的词表示对实体链接结果的影响；语义增强的词表示对实体链接的影响，并对实验结果进行分析。其中还包括对忽略虚词的词表示模型的评测，以及语义增强的词表示模型在词义消歧任务中的验证实验。

第五章是本文的结论和展望部分，总结本文方法中的所取得的成果以及不足，并对未来工作进行展望。

# 2 原理和知识库

## 2.1 词的表示学习

词的表示学习是指通过大规模的语料对词语的语法和语义特征进行自动学习的过程。词的表示学习技术一直受到研究者的关注，近些年有多个词表示的训练模型被提出，并应用于各项自然语言处理任务中，而且在部分任务中获得了当前最高的水平。

### 2.1.1 词的表示学习的发展

词的表示学习最早是Hinton在《Learning Distributed Representations of Concepts》一文中提出的，该文提出了利用神经网络自动生成概念的表示信息[24]。然而由于当时计算机的计算水平以及神经网络的求解算法的限制，并没有得到很好的发展。直到2007年，无监督的受限玻尔兹曼机[25]加入到神经网络，以及反向传播算法[26]对神经网络算法的调优，提高了算法的执行速度。同时硬件的发展例如图形处理器（GPU），显著提高了机器的计算能力，算法的运行时间明显减少。这些条件使得词的表示学习技术得以迅速发展，并被广泛应用。

表 1 主流词表示模型的对比

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **模型** | **目标词与上下文之间的关系** | **上下文的表示方法** | | NNLM[1] | 上下文预测目标词 | 合成 | | LBL[28] | 上下文预测目标词 | 合成 | | C&W[29] | 最大化上下文与目标词联合概率 | 合成 | | Skip-gram[2] | 上下文预测目标词 | 上下文中某个词 | | CBOW[2] | 上下文预测目标词 | 上下文中所有词的平均 | |

目前为止，基本上所有的词表示模型都是基于一个分布式假设：具有相同上下文的词语同时具有相近的语义或语法意义[27]。在此假设之下，人们提出了各种方法对目标词与目标词的上下文之间的关系进行建模。主流的模型在目标词与目标词的上下文之间的关系、目标词的上下文的表示这两个方面有所不同，表1是对当前常用模型的对比。对于目标词与目标词的上下文之间的关系，表1的前四种方法都是使用一个与通过上下文预测目标词的条件概率（*p(w/c)*）相似的目标函数，而C&W模型[29]则是对目标词和上下文之间的联合概率进行最大化。对于目标词的上下文的表示，也是各不相同，或是使用其中一个词进行表示（Skip-gram），或是上下文所有词向量的平均（CBOW），或者上下文词语的组合（LBL,NNLM,C&W）。

此外，除了基于神经网络的方法以外，另一种方法则是基于对上下文的表示矩阵的研究[5,30,31]。其中矩阵的每一行表示一个词语，每一列表示上下文，矩阵中的元素是词语与上下文中的词语的共现次数。

本文选用了当前最流行的CBOW模型、Skip-gram模型和GloVe模型训练词的表示，接下来分别对它们进行详细的介绍。

### 2.1.2 CBOW（Continuous Bag-of-Words Model）

CBOW模型（如图1所示）与上文提到的神经网络语言模型相似，但是隐含层被省略掉，映射层被所有的词语共享，因此所有的词都被映射到同一个位置，对所有词语的表示向量累加，然后平均后的值作为映射层。去掉隐含层之后，模型从神经网络结构转变成log线性结构，由于在计算的时候少了一个矩阵计算，所以大幅度的提升了模型的训练速度。同时该模型类似于词袋模型，所以并没有考虑词语在句子中的顺序。

假设CBOW模型的一段训练样本为*wt-(n-1),…,wt*,输入便为：

|  |  |  |
| --- | --- | --- |
|  |  |  |

模型根据上下文对目标词进行预测：

|  |  |  |
| --- | --- | --- |
|  |  |  |

CBOW模型的优化目标为最大化：

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |
| --- |
| C:\Users\tanggongbo\Desktop\cbow.PNG |

图 1 CBOW 模型结构图

CBOW模型有两种训练方法，分别是分层Softmax方法和负采样方法。由于算法内容不是本文的重点，本文只做简单的介绍。

负采样方法：是NCE(Noise Contrastive Estimation)的一个简化版本。用来提高训练速度和改进词向量质量。该模型每次迭代时都能在增加看到的词语的概率的同时，降低没看到的词语的概率，那么模型会更快的收敛。

分层Softmax方法：通过构造一个树形的结构，降低模型求解的时间复杂度。如果我们有100个数组成的分布，但我们只有他们的非归一化概率，如果想计算出某个数的归一概率，我们需要计算100次。而如果我们把他们分为10类，每类10个数，如果我们知道这10个类的非归一化的概率，以及每类的10个数的非归一化概率，那么我们计算某个数的归一化概率就只要20（10+10）次，这就提高了5倍的速度。同理，进一步分更多的层，利用二分法可提高效率。该方法就是结合哈夫曼编码，由于高频词的隐含层数目要少一些，从而减少计算。输入对应词语的上下文，进行映射，最后输出的哈夫曼树，则其概率是沿着二叉树的所有概率的乘积，最后再进行归一化操作。

### 2.1.3 Skip-gram

Skip-gram模型（如图2所示）与CBOW模型类似，但是与CBOW模型利用上下文预测当前词语的方式相反，Skip-gram模型利用当前词语来预测出现在同一个句子中的其他词语，包括出现在当前词语的前面和后面的所有词语。一般来说，如果一个词语与当前词语的距离越远，那么他们的相关性就越低，所以在该模型的训练过程中，取样的时候就会降低远距离词的权重。训练过程中的优化目标为最大化：

|  |  |  |
| --- | --- | --- |
|  |  |  |

其中，

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |
| --- |
| C:\Users\tanggongbo\Desktop\SKIP-GRAM.PNG |

图 2 Skip-gram模型结构图

Skip-gram模型同样有分层Softmax和负采样两种训练方法，由于在描述CBOW模型时已经介绍过，此处不再赘述。

### 2.1.4 GloVe

与上面提到的CBOW模型和Skip-gram模型基于神经网络的预测机制不同，GloVe模型是通过统计语料中词与词之间的共现信息，通过矩阵分解的技术得到词语的向量表示。虽然很多无监督的词表示学习模型都会用到词语的共现信息，如何利用共现信息对词语进行表示仍然是个问题[5]。

GloVe模型作者通过统计发现，词语之间共现概率的比值更应该应用于词语的表示学习，统计数据如表2所示。相比于stream，solid和ice共现的概率更高，所以概率的比值*P(k|ice)/ P(k|stream)*较大，而gas与stream共现的概率更大，所以概率的比值较小。对于water和fashion，与ice和cream的相关性都类似，因此概率的比值更接近于1[5]。

表 2 词语的共现概率以及比值信息

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **概率和比值** | **k = solid** | **k = gas** | **k = water** | **k = fashion** |
| *P(k|ice)* | 1.9\*10-4 | 6.6\*10-5 | 3.0\*10-3 | 1.7\*10-5 |
| *P(k|stream)* | 2.2\*10-5 | 7.8\*10-4 | 2.2\*10-3 | 1.8\*10-5 |
| *P(k|ice)/ P(k|stream)* | 8.9 | 8.5\*10-2 | 1.36 | 0.96 |

## 2.2 知识库

### 2.2.1 《知网》[[1]](#footnote-1)

《知网》是董振东等提出的以汉语和英语的词语所代表的概念为描述对象，并且将揭示概念之间以及概念所具有的属性之间的关系作为基本内容的常识知识库。《知网》中包含丰富的词语语义知识以及世界知识，为自然语言的各项研究提供了宝贵的资源。本文主要利用了《知网》提供的义原信息，暂未使用世界知识以及概念之间的各种关系。接下来是对《知网》的简单介绍（本文用到的《知网》为2011版）。

《知网》的知识结构中最重要的两个概念是“概念”和“义原”。“概念”可以理解为词语的一个义项，单义词只有一个义项，那么就对应一个“概念”，而一个多义词有多个义项，那么也就对应多个“概念”。“义原”是用来描述“概念”的最小意义单元，所有的“概念”都可以用“义原”进行表示。例如概念“跑酷”的描述语言为：“DEF={fact|事情:CoEvent={exercise|锻炼},domain={sport|体育}}”，其中“事情”、“锻炼”、“体育”均为“跑酷”的描述义原。《知网》并不存在像《同义词词林》和WordNet那样的树状结构，而是通过义原之间存在的关系将所有的概念进行关联，组成一个网状的知识库。《知网》中的义原包含实体、事件、属性、属性值、动态角色与属性、次要特征以及专有名词七大类义原，共计2448个。本文主要利用的就是这些具有语义表示能力的义原知识。

《知网》中以概念为单位进行描述，每一条概念的描述作为一个记录。具体形式如表3所示。其中NO.表示在《知网》中记录的序号，W\_C、G\_C、S\_C、E\_C分别表示中文的词语、词性、情感极性、例子，W\_E、G\_E、S\_E、E\_E分别表示英文的词语、词性、情感极性、例子，而DEF是知网的知识描述语言，是《知网》的核心内容。

表 3 《知网》中概念的描述实例

|  |
| --- |
| NO.=120283  W\_C=少  G\_C=verb [shao3]  S\_C=  E\_C=还~两个杯子，~双筷子，~了三块钱  W\_E=be short  G\_E=verb  S\_E=  E\_E=  DEF={lack|缺少} |

### 2.2.2 维基百科[[2]](#footnote-2)

维基百科是一个免费的在线多语言百科全书，是由全世界的志愿者协作完成。目前维基百科是世界上最大、最流行的网上百科全书，而且是动态、快速增长的。维基百科中最基本的条目是一篇包含定义或者描述一个实体以及话题的文章，而且维基百科中的每篇文章都通过一个唯一的标识符索引。

维基百科中包含大量的命名实体以及与命名实体相关的知识。此外维基百科中大量的结构化知识为命名实体链接提供了有用的特征。例如实体页面、文章目录、重定向页、消歧页以及文章中的超链接。

实体页面用来描述某个特定的实体，一般来说每个页面的标题就是页面描述的实体最常用的名称。重定向页一般是具有相同指向的术语，简写，或者该实体的其他名称的变体。消歧页是用来区分具有相同名称的不同实体，并且包含各个实体对应的链接。可以借助消歧页提取实体的简称或者别名。

## 2.3 语义增强的词的表示学习

### 2.3.1模型

在普通的词表示模型中，一个词语只有一个对应的词向量进行表示，对于多义词来说，一个词向量显然是不够的，Huang等为多义词训练多个词向量[32]，Chen等则为多义词的每个义项训练对应的词向量[7]。因此我们也用不同的词向量来表示多义词的每一个义项，增强词向量的语义表达能力。我们将《知网》中的具有语义表征能力的义原知识融入到词表示的训练过程中，既可以得到普通词语的词向量，又可以得到义原的表示向量，进而通过义原的表示向量来表示词语。图3为在CBOW模型的基础上改进的语义增强的词表示模型结构图。

|  |
| --- |
|  |

图 3 语义增强的词表示模型结构图

### 2.3.2 词的表示

上文在介绍《知网》的时候已经提到词语是由概念构成的，而且多义词由多个概念构成，而每一个概念可以由相应的表示义原进行表示，所以词的表示就转化为概念的表示，概念则是通过义原向量进行表示。对于概念的表示，我们有以下三种方法：

1. sumVec: 用表示概念的义原向量的累加值来表示概念。
2. averVec: 用表示概念的义原向量的平均值来表示概念。
3. allVec: 表示概念的所有义原的向量均用来表示概念。

# 3 搜索查询日志中的实体链接

本文尝试将当前主流的词表示学习的方法应用到查询日志中的实体链接任务中。一方面探究提升词的表示能力的方法，另一方面则是将词的表示学习与传统的实体链接方法结合，探究新的实体链接方法。

词的表示学习方法作为一种无监督的学习方法，不需要人工标注的训练语料就能得到词语的向量表示，大大降低了模型的训练成本，在增加了训练语料规模的同时也避免了在标注语料时引入人为的错误。而经典的词的表示学习方法只利用了原始训练语料中的知识，没有充分利用已有的外部语言学知识，也没有考虑训练语料的优化问题。所以本文会从忽略虚词，以及增加外部语义知识两个方面探究对模型的影响。

对于实体链接中的候选实体模块，本文借鉴传统的候选实体的生成方法，结合基于搜索引擎、维基百科和实体词典的方法生成候选实体。

本文将词表示模型应用到实体链接中最重要的实体消歧模块，首先利用基于词袋模型的相似度计算方法，即计算候选实体的摘要和查询日志之间的相似度进行消歧，由于实体的摘要信息以及查询日志的句子长度有限，包含的特征也就有限，导致此种方法的实验结果并不理想。在此基础上本文引入候选实体本身和实体指称本身之间的相似度，可以明显提升实体链接的准确率。

这一章将具体介绍本文的查询日志中实体链接方法的流程，以及各个模块的具体实现方法。

## 3.1 实体链接流程

本文的实体链接流程如图4所示。整体可以分为四个部分，分别是实验前的数据预处理，词表示模型的探究，候选实体的生成以及实体消歧模块。

数据预处理阶段是对实验所用语料和数据进行分词和词性标注工作。包括为词表示模型提供训练语料，将训练数据和测试数据进行预处理，根据知识库生成实体库，借助维基百科生成同义词词典等。

词表示模型的探究包括三部分。第一部分是借助CBOW、Skip-gram和GloVe模型分别进行训练普通的词表示向量。第二部分是忽略虚词的词表示模型，并将训练结果与普通的词表示模型在词语类推任务中进行比较评测。第三部分则是借助《知网》中的义原知识训练语义增强的词表示模型，并在与实体消歧任务类似的词义消歧任务上进行验证。

本文生成候选实体的方法主要是基于搜索引擎、维基百科以及实体库的方法。首先通过搜索引擎获得实体指称初步的候选实体列表，再通过维基百科和实体库进行过滤，进而得到最终的候选实体。

实体消歧模块也可以分为两部分。对于实体消歧方法，本文根据候选实体和实体指称之间的相似度进行消歧，而相似度的计算方法有两种，一种是基于词袋模型的相似度计算方法，即候选实体的摘要信息与实体指称所在的查询语句之间的相似度。另一种则是改进的基于词袋模型的计算方法，在基于词袋模型的方法基础之上引入候选实体本身与实体指称本身之间的相似度。同时设计将改进的不同的词表示模型应用于实体消歧模块的对照实验。

|  |
| --- |
|  |

图 4 查询日志中实体链接流程图

## 3.2 数据处理

数据处理主要是分词、词性标注、构建实体库词典以及同义词词典。可以为词表示模型提供训练语料；方便候选实体与实体指称之间相似度的计算；过滤候选实体；扩展查询语句。

### 3.2.1 分词与词性标注

中文语料不像英语语料一样有天然的空格可以区分词语的边界，这就需要对中文语料进行分词处理。本文所用的中文分词工具为NLPIR工具包[[3]](#footnote-3)，由于查询语句中有较多的新词，我们将高频新词和实体名加入到分词工具的用户词典中，避免常用的词组以及实体名称被切分。而词性标注是为了区分实词和虚词，去除语料中的虚词，然后利用该语料训练忽略虚词的词表示模型，中文的词性标注工具同样是NLPIR工具包。

评测忽略虚词后的词表示模型时，由于中文领域暂时没有对词表示模型的公开评测数据集，因此我们选择在英文数据集上进行测试，并选择词语类推任务进行测试。 我们使用斯坦福大学的词性标注工具pos-tagger[[4]](#footnote-4)[33]对英文语料进行词性标注。

### 3.2.2 实体词典

本文中实体链接的要求只需要将出现在知识库中的实体进行链接即可，没有出现在知识库中的实体直接忽略，不需要链接操作。而且知识库中包含大量的实体属性信息，为了对候选实体进行筛选，降低实体链接的复杂度，我们将出现在知识库中的所有实体进行汇总并进行一定的处理，得到实体词典。部分实例如表4所示。

表 4 实体词典的构建实例

|  |  |
| --- | --- |
| **原始实体名称** | **新实体名称** |
| “爱情公寓1”，“爱情公寓2”，“爱情公寓3”，“爱情公寓4”，“爱情公寓\_(电视剧)”，“爱情公寓\_(网站)” | “爱情公寓” |
| “爱是永恒”，“爱是永恒（当所爱是你）” | “爱是永恒” |
| “父辈的旗帜\_(书)”“父辈的旗帜\_(电影)” | “父辈的旗帜” |
| “茱莉娅·罗伯茨” | “茱莉娅罗伯茨” |

### 3.2.3 同义词词典

相比于传统文档，搜索引擎中的查询日志语句非常灵活，规范性比较差，含有大量的网络用语，以及昵称、曾用名、简称、别称等。这就会导致实体的识别、候选实体的生成、以及实体消歧模块准确率降低。为了提高候选实体的召回率，实体消歧的准确率，我们利用维基百科和知识库构建同义词词典增加特征。

我们提取知识库和维基百科中出现“绰号”、“别名”、“曾用名”、“简称”、“昵称”、“本名”等标签的信息。例如，“小旋风”是林志颖的昵称；“海贼王”是“ONE\_PIECE”的翻译；“国际田联”是国际田径联合会的简称；谢添原名“谢洪坤”，曾用名“谢俊”；“石将军”是石勇的绰号。同时维基百科中的重定向页也为同义词词典的构建提供同义词资源。例如“爱因斯坦”会重定向到“阿尔伯特·爱因斯坦”；“美国职篮”会重定向到“NBA”。

## 3.3 词表示模型的训练

本文采用基于词表示的方法实现实体的链接，其中最关键的是通过模型训练得到词的表示向量。我们在当前主流的词表示训练模型的基础上，进一步优化和改进。同时得到了忽略虚词后训练得到的表示向量以及融入知网义原知识后训练得到的表示向量。

### 3.3.1 主流的词表示训练

本文选择前面介绍过的当前最主流，效果也最佳的CBOW、Skip-gram和GloVe三个模型训练普通的词表示向量。

三个模型的作者已经开源其训练工具包，本文直接利用工具包word2vec[[5]](#footnote-5)和GloVe[[6]](#footnote-6)将处理好的语料输入到训练模型，设置参数，最终得到词的表示向量。

中英文训练语料分别来自来自中文wikiDumps和英文wikiDumps中的文章数据[[7]](#footnote-7)。其中中文数据相对较小，只有1.13G，而英文数据大概11.2G。由于WikiDumps中的原始数据是XML格式，同时包含各种噪音，经过文本格式转换、中文分词、中文繁体简体转换以及去除标点、符号、数字等操作后中英文训练语料最终分别大概有750M和7.5G。

对于CBOW模型和Skip-gram模型，我们将其向量的维度设置为50、100、200、300、500，选择分层Softmax方法，上下文窗户口为5，同时分配12个线程训练词的表示向量，对出现次数小于5的词语忽略不计。而对于GloVe模型，模型的窗口大小和迭代次数均设置为15。

### 忽略虚词的词表示训练

自然语言中的词语普遍分为实词和虚词两大类，中文和英文都是此类语言。其中实词又可以分为名词、数量词、代词、动词、形容词、副词等；虚词可以分为介词、连词、助词、冠词和语气词等。一般来说，实词能够作为句法成分，而虚词不能够作为句法成分，只能在句子中起某种语法作用[34]。朱德熙在《语法讲义》中也指出实词表示事物、动作、行为、变化、性质、状态、处所、时间等等，虚词只有语法作用，无实际意义，如“的，把，被，所，呢，吗”，或者某种逻辑概念，如“因为，而且，和，或”等等。实词是自由的，即能够单独成句，虚词绝大部分粘着的，即不能单独成句；实词在句法结构中位置不固定，前置，后置都可以，虚词一般都是固定的；实词是开放类，无法一一列举；虚词是封闭类，可穷举[35]。Fries也指出虚词在句子中表达词语之间的语法或者结构关系[36]。因此根据虚词在语言学中的作用可以猜测去掉虚词后的训练语料可能会得到表示能力更强的表示向量。

为了探究忽略虚词后，词的表示能力能否有所改进，我们利用去除虚词后的语料训练另外一份词表示向量。由于中文中没有评价词表示的公开数据集，因此本文选择英文语料进行验证。测试数据来自[2]中词义类推任务所用的数据。该数据集可以分为语法子集和语义子集两部分数据，分别有1万条和9千条问题。语法型问题类似“eat is to eats as go is to (goes)”； 语义型问题类似“Moscow is to Russia as Beijing is to (China)”.验证方法就是判断最接近*vb-va+vc*的向量所表示的词语是否是问题“a is to b as c is to \_?”的答案。

忽略虚词的词表示的训练除了训练语料中虚词被特定的符号替换以外，其他的训练参数同普通的词表示向量的训练，此处不再赘述。中英文语料中去掉的虚词词性标注表如表5所示，其词性标注体系分别为北大一级词性标注[[8]](#footnote-8)和宾州树库词性标注标准[[9]](#footnote-9)。

表 5 虚词词性列表

|  |  |
| --- | --- |
| 中文语料中虚词词性列表 | 英文语料中虚词词性列表 |
| 副词 d  介词 p  连词 c  助词 u  语气词 y | CC(Coordinating Conjunction)  EX(Existential there)  IN(Preposition/subordinating conjunction)  MD(Modal)  RP(Particle)  TO(to)  WDT(Wh-determiner)  WP(Wh-pronoun)  WP$(Possessive Wh-pronoun)  WRB(Wh-adverb) |

### 3.3.3 语义增强的词表示训练

《知网》中的义原是表示概念语义的基本单位，因此具有语义表示能力，我们将义原知识融入到普通的词表示训练模型中，目的在于增强词表示向量的语义表示能力。

《知网》中，如果一个词语只有一个中文概念，而且概念只用一个义原进行描述，那么我们将这个词定义为单义原词，反之则定义为多义原词。经过统计发现，单义原词共有35347个，构成单义原词的义原数量为1492，约占义原总数的60.95%；多义原词共有69382个，构成多义原词的义原数量为2041，大约占义原总数的83.37%。其中多义原词中71.88%（1467/2041）的义原也出现在单义原词中的义原（见图5），这就说明大部分的词语是可以通过单义原词中的义原来进行语义的表示。同时，平均每个单义原词中的义原可以表示23.7个单义原词，而且大部分的单义原词由少量义原构成（见图6），说明利用义原来表示单义原词可以明显减少特征的训练。此外，由图7可以发现，绝大多数的多义原词由少于7个义原构成，说明利用义原向量来表示多义词也是简单可行的。与此同时,语料中的多义词有多个表示，这就提高了词表示的语义表示能力。

词表示模型的训练过程可以分为以下步骤：

1. 选取已经处理好的中文训练语料。
2. 将语料中出现的单义原词替换为对应的义原，同时加上标签，比如：“参观”被替换为“【看】”，与原始语料一同作为训练数据。通过训练我们就可以得到表示词语“参观”和义原“看”的向量。由于单义原词中义原数量远远小于单义原词本身（4%），因而这一过程可以大大减少数据稀疏对模型造成的影响。
3. 将处理后的语料输入到训练模型中，并设置相应的参数。

|  |
| --- |
|  |

图 5 义原的分布信息

|  |
| --- |
|  |

图 6 单义原词中的义原表示能力

|  |
| --- |
|  |

图 7 多义原词中的义原个数统计图

为了验证这种词表示向量的有效性，本文将其应用于与实体链接相似的词义消歧任务中，其中，义原和词语的表示向量通过第二章中提到的语义增强的词表示模型训练所得，模型的其他参数也同第一节中词表示的训练模型。我们选择SENSEVAL-3中的中文词义消歧任务，其中包含20个汉语词的379个实例。例如“材料”这个多义词的消歧实例如图8所示。

|  |
| --- |
|  |

图 8 “材料”的消歧实例

本文主要通过计算多义词与多义词上下文的特征向量之间的语义相似度来对多义词进行消歧，相似度最高的义项被认为是多义词在该语境下的语义。

多义词与特征向量之间的相似度计算主要通过计算多义词的概念与多义词的特征向量之间的相似度来实现。选取相似度最大的概念作为该多义词在当前上下文的解释。假设多义词有N个概念，那么概念向量集合为{ c1，c2 ，……，cn}。假设特征向量设为F。则有：

|  |  |  |
| --- | --- | --- |
|  |  |  |

C＇即为多义词在当前上下文的最佳的候选义项（概念）。而*sim(ci，F)*表示概念与特征向量的相似度。

前文提到过，概念一般是由多个义原进行表示，其表示方法有：

a）sumVec：将表示概念的义原向量进行累加，累加结果用来表示概念。

b）averVec：求表示概念的义原向量的平均值，平均值用来表示概念。

c）allVec：表示概念的所有义原向量均用来表示概念。

而计算概念与特征之间的相似度也有两种方法：

1. 使用sumVec和averVec方法表示向量时，相似度为两个向量的余弦距离。

b）使用allVec表示向量时，采用Mihalcea的向量对齐的方法[37]计算相似度。

向量对齐方法：假设有J个向量表示特征，特征向量集合为F：{ *f*1，*f*2 ，……，*f*j }，有K个向量表示概念，概念向量集合为C：{ *c*1，*c*2 ，……，*c*k }，F和C中的向量两两之间计算余弦相似度，取相似度最大的一对作为已对齐的向量，插入到集合P中。*P：*{{(*f*l，*c*m)……}（*f*l∈*F*，*c*m∈*C*）。然后从F和C中将已选择的向量删除，循环执行此步骤，直至F或者C变成空。最后计算所有对齐向量的平均余弦相似度：

|  |  |  |
| --- | --- | --- |
|  |  |  |

其中，n是j和k中较小的一个数。

|  |  |  |
| --- | --- | --- |
|  |  |  |

本文选取多义词所在的上下文作为特征。Niu等认为词语的上下文窗口为10以内的时候，实验结果最好[38]。Ke的实验结果表明上下文特征窗口为5（前后均为5）的时候准确率最高[39]。Huang的实验结果则表明上下文窗口为1的时候效果最好[40]。本文分别选取上下文窗口为1的实词和待消歧语句中除多义词以外的所有实词作为特征，设置了两组实验。由于选取的上下文特征中一般都不仅仅只有一个词语，所以如何表示特征向量又是一个问题。针对这个问题，我们采用了与概念的表示相同的三种方法：

1. sumVec：上下文窗口内实词向量的累加值作为特征向量
2. averVec：上下文窗口内实词向量的平均值作为特征向量
3. allVec：上下文窗口内所有实词向量均作为特征向量

## 3.4 候选实体的生成

本文利用基于搜索引擎、维基百科、以及实体词典的方法生成候选实体。

通过与实体词典的匹配，我们将实体指称分为两类，一类是可以与实体词典中的实体完全匹配的实体指称，我们称之为确定的实体指称*Mi*，例如“爱情公寓”，“山东大学”，反之则成为非确定的实体指称*Mu*，例如“新华人寿”，“湖人”等。由于查询日志并非规范文本，实体指称中存在大量的噪音，而同义词词典可以解决一部分实体指称的变形，但是像在查询日志中出现的“沙糖桔”，在知识库中对应的却是“沙糖橘”，在我们的同义词词典中却没有该实体的别名信息，因此只利用词典是不能够得到所要链接的实体。而搜索引擎能够对查询日志进行一定的文本规范处理，在搜索前能够自动将“沙糖桔”处理成“沙糖橘”，所以我们同时采用基于搜索引擎的候选实体生成方法，同时加入维基百科的知识，利用搜索引擎百度，搜索实体指称+“中文维基百科”来获取候选实体，并进一步利用实体词典对候选实体进行过滤。实验流程如图9所示。

|  |
| --- |
|  |

图 9 候选实体的生成图

对于确定的实体指称*Mi*，而且知识库中只有一个对应的实体，如果第一个搜索结果是中文维基百科页面，有以下三种情况：

1. 该页面是一个确定页面时，标题中的实体便是我们要链接到的实体，例如“诺基亚”。
2. 该页面是重定向页，标题中的实体也是我们要链接到的实体。
3. 页面是一个消歧页，将页面中所有实体作为候选实体，进入下一步候选实体的排序，例如：实体指称“杨颖”的候选实体有“Angelababy”和“杨颖（作家）”。

对于确定的实体指称*Mi*，而知识库中对应多个实体，那么这些对应的实体则作为候选实体。例如“倚天屠龙记”，在知识库中对应多个实体：“倚天屠龙记”，“倚天屠龙记\_(1978年电影)”，“倚天屠龙记\_(1978年电视剧)”，“倚天屠龙记\_(1986年电视剧)”，“倚天屠龙记\_(1994年电视剧)”，“倚天屠龙记\_(2001年电视剧)”，“倚天屠龙记\_(2003年电视剧)”，“倚天屠龙记\_(2009年电视剧)”。这八个实体均作为候选实体进行后续的候选实体的排序。

对于不确定的实体指称*Mu*，如果前10条搜索结果中有中文维基百科页面，那么将其标题中的实体均作为候选实体，进入下一步实体消歧模块。由于评测任务只需要将知识库中有的实体进行链接，如果前10条搜索结果中没有中文维基百科页面，那么我们直接将该实体指称作删除处理。同时，利用实体词典对候选实体进行过滤，减小实体消歧的难度。

## 3.5 实体消歧

本文首先计算候选实体和实体指称之间的余弦距离作为相似度，然后将相似度的值输入到分类器中进行分类，最终得到实体链接的结果。对实体进行消歧，相似度的计算方法也分为两种，一种是最基本的基于词袋模型的计算方法，用候选实体的摘要信息和实体指称所在的查询日志之间的相似度进行表示；另一种方法是在第一种基于词袋模型的方法的基础上，通过观察和分析实验结果后，增加了候选实体本身和实体指称本身之间的相似度，用来提高实体消歧的准确率。

由于是基于词表示的实体链接，在实体消歧模块中，所有的相似度都是通过词的表示向量计算所得。同时，我们将前面训练所得的所有词表示模型均应用于实体消歧模块中，探究不同的模型对实体链接结果的影响。

### 3.5.1 实体指称与候选实体之间的相似度

1. 基于词袋模型的相似度

词袋模型是指将一段文本或者文档看作是一袋子的词语，不去考虑语法和词序的关系，每个词都是独立的。本文中，我们分别计算查询日志语句和候选实体的摘要信息中所有词表示向量的平均值，并分别用来表示查询日志和候选实体的摘要信息。那么基于词袋模型的相似度*Simbow*计算方法如公式3-4所示。*C*表示当前实体指称*m*对应的候选实体，*Nq*表示查询日志中词语的个数，*Na*表示候选实体的摘要信息中词语的个数，*Vqi*和*Vaj*分别表示查询日志和候选实体的摘要信息中词语的向量表示。

|  |  |  |
| --- | --- | --- |
|  |  |  |

1. 改进的基于词袋模型的相似度

传统的词袋模型通常应用于长文本文档，而上面的方法中所用到的查询日志和候选实体的摘要文本普遍非常短，只有几十个词，甚至只有十个词左右，单独依靠词袋模型是不能够很好的表示查询日志语句和候选实体的摘要信息，同时也不能很好地得到用于实体消歧的关键特征。通过分析实验结果和实验数据，我们在上面的方法的基础上引入了候选实体本身和实体指称本身之间的相似度，即计算这两个词语之间的相似度。那么改进的相似度*Simibow*的计算方法如公式3-5所示。其中*vc*表示候选实体的向量表示，*vm*则是实体指称的向量表示。

|  |  |  |
| --- | --- | --- |
|  |  |  |

### 3.5.2分类器的训练

判断实体指称的候选实体是否是正确的链接实体可以看作是一种二分类的问题，本文借助LIBSVM工具包[[10]](#footnote-10)[41]中的SVM算法进行分类。

首先是将训练数据中的实体指称和对应的候选实体之间的相似度值，以及链接结果输入到模型中，通过训练得到分类模型的参数，然后再将测试数据中的实体指称和候选实体之间的相似度的值输入到已经训练好的模型中，最终得到测试数据的实体链接结果。

### 3.5.3词表示模型与实体消歧

为了探究改进的词表示模型对实体链接结果的影响，本节介绍我们设置的两组对照实验，分别是忽略虚词的词表示模型与普通词表示模型的对照实验，以及语义增强的词表示模型与普通词向量的对照实验，并且均采用改进的基于词袋模型的相似度计算方法。

# 4 实验

本章将介绍实体链接的评价方法、对忽略虚词的词表示模型的评测、以及语义增强的词表示模型在词义消歧任务中的测试、实体消歧中不同相似度的计算方法的对比、不同词表示模型在实体链接中的表现，以及对实验结果的分析。

## 4.1 评价准则

本文所用的训练数据以及测试数据均来自2015年自然语言处理与文字计算中的“Entity Recognition and Linking in Search Queries”评测任务，所以我们采用相同的评价准则。

该任务要求识别出查询日志中出现的所有实体，并将其链接到给定的知识库，如果知识库中不存在相应的实体，则不作链接处理。例如，给定查询日志：“射雕英雄传刘亦菲版”，目标就是将“射雕英雄传”链接到知识库中的“射雕英雄传（2006年电视剧）”，同时将“刘亦菲”链接到知识库中的“刘亦菲”。而对于查询日志“射雕英雄传”，其中的“射雕英雄传”是一个有歧义的实体，可能是指的一本书，可能是某一个版本的电视剧，此时需要链接的实体就不仅仅是一个确定的实体，而是“射雕英雄传”、“射雕英雄传（1983年电视剧）”、“射雕英雄传（1994年电视剧）”、“射雕英雄传（2003年电视剧）”、“射雕英雄传（2006年电视剧）”等。

对于一条查询日志，实体链接系统*S\**的输出中共有| *S\**|组不同的链接结果，其中被识别而且进行链接的组数|*Sr\**|，被识别而且链接正确的组数为|*Sc\**|，人工标注系统S共标注|*S*|组不同的链接结果。设链接准确率为*Link-Precision*，链接召回率为*Link-Recall*，链接F值为*Link-F1*，那么则有：

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

如果共有n条查询日志，设平均F值为Average-F1，则:

|  |  |  |
| --- | --- | --- |
|  |  |  |

本文主要是是对实体链接的研究，因此实体识别部分将不再考虑，我们直接对已经识别的实体进行链接工作。

## 4.2 实验数据

本文所用的训练数据以及测试数据均来自2015年自然语言处理与文字计算中的“Entity Recognition and Linking in Search Queries”评测任务。其中训练数据中有159条已经标注好的查询日志，包括识别的实体指称以及正确的链接实体。而测试数据则是由3849条真实的查询日志语句组成。此外还有一个给定的知识库，用来链接实体，知识库是以三元组的形式存在，包括实体的名称以及各种属性，属性值等。训练数据，测试数据以及知识库的样例如表6所示。

表 6 实验数据样例

|  |  |
| --- | --- |
| **数据类别** | **实例** |
| 训练数据 | 神雕侠侣黄晓明版 神雕侠侣 62218 |
| 测试数据 | 54 广州恒大 亚冠 |
| 知识库 | 54722 0 0 孙中山 abstract '孙文……国民党总理。'  54722 0 0 孙中山 name 孙文  ……  54722 0 130533 孙中山 country 北洋政府 |

对于普通词表示模型的训练，中英文训练语料分别来自中文wikiDumps和英文wikiDumps中的文章数据。训练数据的统计信息如表7所示。经过数据清洗与处理后，中英文训练语料最终分别大概为750M和7.5G。英文语料中，我们只统计出现次数大于5的词，词表大小约为198万，词的总数约为21.6亿。 对于中文语料，由于语料本身比较小，而且语料的清洗和处理比较彻底，同样只统计出现次数大于5的词，词表大小只有17.6万，词的总数大约为1.5亿。对于忽略虚词的词表示模型，去掉前文中给定词性的虚词后，中文语料的词表大小约为17.5万，词的总数变为1.28亿；英文语料的词表大小变为195万，词的总数变为21.3亿。对于语义增强的词表示模型，替换训练语料中的单义原词后，得到约1.8万个义原实例，词数约为3亿。

表 7 词表示模型中训练语料的统计信息

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **语料** | **词表大小** | **词数** |
| 普通词表示模型 | 中文 | 17.6万 | 1.5亿 |
| 英文 | 198万 | 21.6亿 |
| 忽略虚词的词表示模型 | 中文 | 17.5万 | 1.28亿 |
| 英文 | 195万 | 21.3亿 |

## 4.3词表示与实体链接

在第三章中我们已经提到过本文借助词的表示向量，采用了两种计算相似度的方法对实体进行消歧，一种是基于词袋模型相似度，另外一种是改进的基于词袋模型的相似度。两种方法的实验结果如表8所示，其中所有向量的维度均为200。

表 8 不同的相似度计算方法以及评测系统的对比实验

|  |  |  |
| --- | --- | --- |
| **方法** | **词表示训练模型** | **Average-F1** |
| 基于词袋模型的方法 | CBOW | 0.331 |
| Skip-gram | 0.365 |
| GloVe | 0.367 |
| 改进的基于词袋模型的方法 | CBOW | 0.547 |
| Skip-gram | 0.592 |
| GloVe | 0.590 |
| 评测任务第一名 | - | 0.733 |

由实验结果可以发现，改进的基于词袋模型的相似度计算方法获得的结果要明显高于基于词袋模型的相似度计算方法。说明候选实体本身和实体指称本身之间的相似度更有利于实体的消歧。

通过分析实验结果以及实验数据，发现有以下几点导致两个方法之间存在明显的差距：

1. 候选实体的摘要信息和实体指称所在的查询日志的文本长度太短，尤其是查询日志的长度，平均只有十个词左右，所以很难通过词表示向量来对其进行表示，进而导致两者之间的余弦相似度比较低，无法很好地对实体进行消歧。
2. 每一条查询日志中不一定只有一个实体指称，而且不同的实体指称之间的跨度可能会比较大，所以针对同一个查询日志中的不同的实体指称，使用同一个查询日志的表示是不合理的。
3. 候选实体本身和实体指称本身之间的相似度在某种程度上包含了两个词语的字符相似度，对于正确的候选实体，两者之间余弦距离会很大，进而使得候选实体和实体指称之间的相似度比较高，最后成为目标链接实体。而对于错误的候选实体，恰恰相反，余弦距离比较小，也会很容易判断正确。

同时，实验最终的平均F值相比当时评测第一名系统的平均F值要低一些，原因可以总结为以下几条：

1. 本文基于搜索引擎和词典的候选实体生成方法不能够保证很高的召回率。例如：网络商城“天猫”的候选实体为“淘宝网”和“猫座”，并没有“天猫”。
2. 查询日志文本的不规范性同样会降低候选实体的召回率，同时，构建的同义词词典不能够覆盖所有的非规范用语的实体。例如：“软考”是“全国计算机技术与软件专业技术资格（水平）考试”的简称，但是我们的同义词词典中并没有对应的信息；还包括外国人名的翻译问题，“罗曼雷恩斯”在知识库中对应的是“罗曼·瑞恩斯”，因此得到的候选实体中不存在正确的目标实体。
3. 与英文不同，中文语料没有天然的分词符号，因此处理中文语料需要分词操作。对于处理实体相关的任务，分词时保证实体被正确的切分是很困难的，即使在分词的时候加入用户自定义的词典。同时存在数据稀疏的问题，因此部分实体并没有相应的词向量表示，降低了实体消歧模块中准确率。例如：“阿拉斯加幼犬”对应的候选实体是“阿拉斯加雪橇犬”和“阿拉斯加州”，在词表示模型的训练语料中出现的“阿拉斯加雪橇犬”，一般经过分词后的结果是“阿拉斯加”、“雪橇”和“犬”，因此没有对应的表示向量，计算候选实体和实体指称之间的相似度的方法变为基于词袋模型的方法。

## 4.4 忽略虚词的词表示与实体链接

本节将介绍忽略虚词后的词表示的评测实验，以及在实体链接上的对比实验，并对实验结果进行分析和总结。

### 4.4.1 忽略虚词对词表示性能的影响

本文采用当前最主流的三个词表示训练模型（CBOW，Skip-gram，GloVe）训练词的表示向量，并且设置不同的维度（50，100，200，300，500），所有的数据均是英文文本。在词语类推任务中的实验结果如表9所示，单位为百分比，加黑的分数表明忽略虚词后准确率有所上升，而带下划线的分数则表示忽略虚词后准确率有所下降。

表 9 词表示模型在词语类推任务中的对比实验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型** | **维度** | **词语类推** | | | |
| **语义任务** | | **语法任务** | |
| **有虚词** | **无虚词** | **有虚词** | **无虚词** |
| **CBOW** | 50 | 35.38 | **38.45** | 29.62 | 27.24 |
| 100 | 44.33 | **48.80** | 38.20 | 37.49 |
| 200 | 49.59 | **56.21** | 48.08 | 47.63 |
| 300 | 53.84 | **56.31** | 51.69 | 49.62 |
| 500 | 53.68 | **55.28** | 53.11 | 51.52 |
| **Skip**-**gram** | 50 | 27.01 | **42.09** | 40.87 | 25.14 |
| 100 | 53.03 | **56.68** | 38.65 | 36.22 |
| 200 | 64.67 | **65.78** | 45.69 | 43.31 |
| 300 | 67.81 | **68.44** | 49.84 | 47.23 |
| 500 | 69.75 | 69.31 | 52.06 | 50.04 |
| **GloVe** | 50 | 54.39 | 53.07 | 34.75 | **35.03** |
| 100 | 67.46 | 66.08 | 46.40 | **46.99** |
| 200 | 74.34 | 74.08 | 54.29 | **55.04** |
| 300 | 76.71 | 76.00 | 56.67 | **57.88** |
| 500 | 78.49 | 78.35 | 57.74 | **58.20** |

由实验结果可以发现，对于语义型子集，忽略虚词后，除了一个例外的（Skip-gram训练的500维的模型）准确率有微小的下降，其他通过CBOW和Skip-gram训练出来的所有词表示模型的准确率都有所提升。可以理解为由于将虚词替换成特定符号后，模型训练过程中的词表变小，因此能够更好地获得词语的语义特征。而对于语法型子集，结果恰恰相反，所有通过CBOW和Skip-gram训练出来的所有词表示模型的准确率都是降低的。这种现象可以认为是在模型的训练过程中忽略了在句子中具有语法和结构功能的虚词，导致词表示中的语法特征变少。但是对于GloVe模型，结果却与CBOW和Skip-gram模型完全相反，语法任务的准确率是提升的，而语义任务的准确率却是下降的。

图10和图11是三个模型在忽略虚词后训练得到的词表示在词语类推任务中准确率的影响，可以发现忽略虚词对词表示模型在词语类推任务上的影响是最小的，包括语义和语法子集。我们初步认为是忽略虚词在GloVe模型上的影响是可以忽略不计的。同时，也可以看出由Skip-gram模型训练的50维的词表示模型特别突出，应该是一个特例。

根据上面的实验结果，本文认为忽略训练语料中的虚词后得到的词表示模型会提高语义任务上的准确率，而在语法任务上，准确率会有所牺牲。

|  |
| --- |
|  |

图 10 忽略虚词后词表示模型在语义子集任务中准确率的变化

|  |
| --- |
|  |

图 11 忽略虚词后词表示模型在语法子集任务中准确率的变化

### 4.4.2 基于无虚词的词表示的实验与分析

表10是忽略虚词的词表示模型和普通词表示模型的对比实验结果，采用改进的基于词袋模型的相似度计算方法。

表 10 不同相似度计算方法的实体链接结果

|  |  |  |
| --- | --- | --- |
| **训练模型** | **Average-F1** | |
| **普通词表示模型** | **忽略虚词的词表示模型** |
| CBOW | 0.547 | 0.554 |
| Skip-gram | 0.592 | 0.596 |
| GloVe | 0.590 | 0.590 |

由表10可以看出忽略虚词后对实验结果有轻微的提升，但是并不明显。我们总结有以下原因：

1. 由于是基于词袋模型的方法计算相似度，因此表示候选实体的摘要信息和实体指称所在的查询日志的时候不再仅仅是语义或者语法单方面的构成，而是均有混合，只是语义的成分更多一些。因此虽然在语义方面会有提升，但同时受语法因素的影响，导致提升很小。
2. 相比于普通的词汇，候选实体以及实体指称在语料中出现的频次要明显低一些，因此忽略虚词后对其影响也会更小。

## 4.5 语义增强的词表示与实体链接

本节首先对词义消歧实验进行分析，判断语义增强的词表示模型是否是有效的。然后是语义增强的词表示模型和普通词表示模型在实体链接中的对比实验与分析。

### 4.5.1 语义增强的词表示在词义消歧任务中的实验

利用语义增强的词表示模型，根据不同的向量表示和特征表示方法，我们设置了六个对比实验，在SENSEVAL-3中的中文词义消歧任务中的实验结果如表11所示，详细实验结果见表12，其中词向量的维度设置为200。

表 11 词义消歧任务中的对比实验

|  |  |  |  |
| --- | --- | --- | --- |
| **向量表示方法征选取**  **特征选取** | **sumVec** | **averVec** | **allVec** |
| **最近邻实词** | 实验一(29.6%) | 实验三(31.1%) | 实验五(26.1%) |
| **上下文所有实词** | 实验二(32.7%) | 实验四(37.7%) | 实验六(25.6%） |

表 12 词义消歧的实验结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **词** | **义项数** | **实例数** | **实验（正确数）** | | | | | |
| **一** | **二** | **三** | **四** | **五** | **六** |
| 路 | 9 | 28 | 6 | 6 | 6 | 8 | 8 | 6 |
| 冲击 | 4 | 13 | 4 | 7 | 4 | 8 | 6 | 6 |
| 运动 | 5 | 27 | 6 | 9 | 6 | 9 | 5 | 6 |
| 日子 | 3 | 21 | 13 | 11 | 13 | 14 | 10 | 9 |
| 老 | 12 | 26 | 4 | 5 | 4 | 5 | 3 | 4 |
| 没有 | 4 | 15 | 4 | 5 | 4 | 7 | 7 | 8 |
| 穿 | 5 | 14 | 3 | 2 | 3 | 2 | 4 | 2 |
| 地方 | 7 | 17 | 6 | 2 | 6 | 5 | 0 | 0 |
| 活动 | 7 | 16 | 7 | 8 | 7 | 8 | 4 | 6 |
| 少 | 4 | 19 | 8 | 9 | 8 | 10 | 7 | 11 |
| 坐 | 6 | 12 | 1 | 2 | 1 | 5 | 2 | 6 |
| 分子 | 3 | 16 | 5 | 2 | 5 | 6 | 6 | 1 |
| 把握 | 4 | 15 | 8 | 7 | 8 | 7 | 0 | 0 |
| 突出 | 6 | 15 | 6 | 6 | 6 | 6 | 0 | 0 |
| 走 | 9 | 24 | 8 | 8 | 8 | 8 | 7 | 7 |
| 包 | 11 | 36 | 2 | 6 | 5 | 6 | 5 | 2 |
| 起来 | 6 | 20 | 2 | 3 | 3 | 3 | 3 | 0 |
| 钱 | 5 | 20 | 5 | 7 | 6 | 7 | 5 | 7 |
| 研究 | 2 | 15 | 7 | 11 | 7 | 11 | 11 | 11 |
| 材料 | 3 | 10 | 7 | 8 | 7 | 8 | 6 | 5 |
| **总数** | 115 | 379 | 112 | 124 | 118 | 143 | 99 | 97 |
| **正确率** | | | 29.6% | 32.7% | 31.1% | 37.7% | 26.1% | 25.6% |

在相同的测试数据上，我们的实验结果要稍高于Li等将词语的搭配信息和主题信息作为特征，利用贝叶斯分类器实现词义消歧的无监督的方法[42]，但是我们的方法更加简单，不需要复杂的特征工程的工作。

消歧的准确率相比有监督的方法还是有比较大的差距，SENSEVAL-3相应的评测中，最优的系统准确率可以达到66.5%。一方面是由于有监督的方法中测试数据和训练数据的相似度比较高，降低了消歧的难度；另一方面，我们的实验所用的《知网》版本为2011版，给定多义词的义项数明显多于当年的数量，约为评测时2002版的《知网》提供的义项数的146%，增加了我们的难度。此外，还有数据稀疏导致的部分词没有对应的词表示向量，以及个别义项的表示义原比较接近，义项（概念）之间的区分度不够明显，导致无法对多义词正确的消歧。如表13所示，“钱”的部分消歧结果与分析。

表 13 “钱”的部分消歧结果与分析

|  |  |  |  |
| --- | --- | --- | --- |
| **消歧句** | **正确义项** | **消歧结果** | **分析** |
| 丁佩玲（女）、……顾煜麟、<head>钱</head>月宝（女）、钱红菱（女）、钱易（女）、钱海鑫、徐家基、徐德郁（女）…… | 姓 | 无 | 消歧句的上下文完全由人名构成，由于数据稀疏的关系导致没有对应的词向量，无法进行消歧。 |
| 她又用这笔<head>钱</head>买了一头耕牛，发展生产，效果显著。 | 资金 | 资金 | 消歧正确 |
| 临走，李清林留下５００元<head>钱</head>，说这是个人的一点心意，暂时解决一下困难。 | 货币 | 资金 | 由于表示货币和资金的义原比较相似，未能够区别。  货币：货币、商业、金融  资金：资金、金融 |

### 4.5.2 基于语义增强的词表示的实体链接实验与分析

由于语义增强的词表示模型是在CBOW模型的基础上做的改进，因此对照实验是CBOW模型和语义增强的词表示模型两个模型之间的对比，同样采用改进的基于词袋模型的相似度计算方法。

表 14 普通词表示模型与语义增强的词表示模型的对比实验

|  |  |  |
| --- | --- | --- |
| **训练模型** | **表示向量维度** | **Average-F1** |
| CBOW | 200 | 0.547 |
| 语义增强的词表示模型 | 200 | 0.560 |

由表14可以看出语义增强的词表示模型应用于实体链接后，平均F值是有一定的提升，可以理解为将知网中的义原知识融入到词表示的训练模型后可以提升词向量的语义表示能力，并同样适用于实体链接任务。但是同时也发现平均F值的变化只有1.3%左右，并不是特别明显。造成这种结果的原因主要是：

1. 实体消歧过程中，并没有对候选实体摘要和查询日志中出现的多义词进行消歧，与前面提到的词义消歧中表示多义词的方法不同，并不是通过义原表示多义词的某个义项来表示多义词，而是直接利用的多义词的词表示，某种程度上并没有受益于语义增强的词表示模型。因此在计算候选实体与实体指称之间的相似度的时候影响也不会太大。
2. 对于实体，种类和内容千变万化，不会像多义词一样在《知网》等知识库中存在完整的语义信息，尤其是像普通的人名和地名，更不可能出现在《知网》中，因此对于这类实体的影响也基本可以忽略。例如查询日志“玉素甫·哈斯·哈吉甫”只是一个人名，无法通过义原进行表示。

# 5 总结和展望

## 5.1 总结

本文主要介绍了基于词表示的查询日志中的实体链接方法。一方面是对词表示模型的探究与改进，另一方面是对实体链接方法的研究，然后将词表示技术应用到查询日志中实体链接的任务中。对于候选实体的生成模块，我们采用基于搜索引擎和词典的方法；对于实体的消歧，利用候选实体和实体指称之间的相似度进行消歧，其中包含基于词袋模型的相似度计算方法和改进的基于词袋模型的相似度计算方法。同时将普通的词表示模型和改进的词表示模型（忽略虚词的词表示模型和语义增强的词表示模型）应用于实体链接。

通过实验发现：忽略虚词后，可以使得词表示模型在语义相关的任务上有更高的准确率，而会降低模型在语法相关的任务上的准确率；将《知网》中的义原知识融入词表示的训练模型中会提升模型的语义表示能力，并在中文词义消歧任务中得到验证；相比于基于词袋模型的相似度计算方法，加入候选实体本身和实体指称本身之间的相似度后，实体链接的平均F值有较大的提升，从0.365提升到0.592。同时，应用忽略虚词的词表示模型和语义增强的词表示模型后，实体链接的平均F值分别由0.592、0.547提升到0.596和0.560。

## 5.2 未来工作

基于本文词表示模型与实体链接方法的研究，未来的工作主要可以围绕以下几个方面展开：词表示模型的优化与改进；知识库等外部知识的进一步挖掘；实体链接方法的改进。

对于词表示模型，本文只是对已有的词表示模型进行了简单改造，模型的复杂度和准确性方面比较欠缺，导致无法精确地对所有的词语进行表示。而当前各种深度神经网络技术已经应用于自然语言处理的各项任务中，并且取得了当前的最优结果，例如深度神经网络（Deep Neural Networks）、长短时记忆模型的循环神经网络(LSTM RNN: Long-Short Term Memory Recurrent Neural Network)、注意力模型（Attention Model）等等，因此可以尝试将这些模型应用于词表示的训练过程中。此外，可以对实体进行建模，训练针对实体的表示向量，这样就可以更准确的计算候选实体与实体指称之间的相似度，进而提高实体链接的准确率。

本文只是利用了维基百科中页面的属性信息以及《知网》中的义原信息，而这些内容相对于目前已有的知识库来说只是冰山一角，可挖掘利用的信息还有很多。首先是知识库的范围上，我们可以拓展到《同义词词林》以及WordNet[[11]](#footnote-11)、FrameNet[[12]](#footnote-12)、BabelNet[[13]](#footnote-13)等对于外部知识库的挖掘；对于内容上，我们可以挖掘其中的词汇或者实体之间的各种关系，包括上下位关系、同义关系、反义关系、整体与部件、施事和受事关系、事件与角色关系等等。

此外，本文使用的实体链接方法只是在词袋模型的基础上进行一些简单的改进，方法比较简单，所以需要进一步地提出更优的方法。例如：同一条查询日志中可能不仅仅存在一个实体，而且实体之间一般存在联系，因此可以考虑对出现在同一个查询日志中的实体进行协同链接；由于实体链接的知识库在实体链接的过程中很重要，因此我们可以借助当前热门的知识图谱的相关技术对知识库进行表示，来提高实体链接的准确率；最后，由于词义消歧和实体链接中的实体消歧模块都是消歧任务，而且本文中利用语义增强的词表示模型时并没有对多义词进行表示，因此可以采取词义消歧和实体链接联合消歧的方法提高实体链接的准确率。

# 参考文献

1. Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.
2. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
3. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
4. Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]//HLT-NAACL. 2013: 746-751.
5. Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]//EMNLP. 2014, 14: 1532-1543.
6. Yin W, Schütze H. Learning Word Meta-Embeddings by Using Ensembles of Embedding Sets[J]. arXiv preprint arXiv:1508.04257, 2015.
7. Yu M, Dredze M. Improving Lexical Embeddings with Semantic Knowledge[C]//ACL (2). 2014: 545-550.
8. Johansson R, Pina L N. Embedding a semantic network in a word space[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1428-1433.
9. Chen X, Liu Z, Sun M. A Unified Model for Word Sense Representation and Disambiguation[C]//EMNLP. 2014: 1025-1035.
10. Ling W, Chu-Cheng L, Tsvetkov Y, et al. Not all contexts are created equal: Better word representations with variable attention[J]. 2015.
11. Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons[J]. arXiv preprint arXiv:1411.4166, 2014.
12. Dredze M, Mcnamee P, Rao D, et al. Entity Disambiguation for Knowledge Base Population[J]. International Conference on Computational Linguistics, 2010, (3):277-285.
13. Han X, Zhao J. 2009. Nlpr-kbp in tac 2009 kbp track: A two-stage method to entity linking[J]. Proceedings of Test Analysis Conference, 2009.
14. Meng Z, Yu D, Xun E. Chinese Microblog Entity Linking System Combining Wikipedia and Search Engine Retrieval Results[J]. Communications in Computer & Information Science, 2014.
15. Zheng Z, Li F, Huang M, et al. Learning to link entities with knowledge base[J]. Proceedings of the Annual Conference of the North American Chapter of the Acl, 2010.
16. Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 782-792.
17. Zhang W, Yan C S, Su J, et al. NUS-I2R: Learning a Combined System for Entity Linking[J]. Proc Tac Workshop, 2010.
18. Shen W, Wang J, Luo P, et al. LINDEN: linking named entities with knowledge base via semantic knowledge[C]// Proceedings of the 21st international conference on World Wide Web. ACM, 2012.
19. Kulkarni S, Singh A, Ramakrishnan G, et al. Collective Annotation of Wikipedia Entities in Web Text[J]. Assoc Computing Machinery, 2009:457--466.
20. Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 68-76.
21. Wang Z, Zhang J, Feng J, et al. Knowledge Graph and Text Jointly Embedding[C]//EMNLP. 2014: 1591-1601.
22. He Z, Liu S, Li M, et al. Learning Entity Representation for Entity Disambiguation[C]//ACL (2). 2013: 30-34.
23. Ling X, Singh S, Weld D S. Context representation for named entity linking[C]//Proceedings of the 3rd Pacific Northwest Regional NLP Workshop. 2014.
24. Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.
25. Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 791-798.
26. Hecht-Nielsen R. Theory of the backpropagation neural network[C]//Neural Networks, 1989. IJCNN. International Joint Conference on. IEEE, 1989: 593-605.
27. Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.
28. Mnih A, Hinton G. Three new graphical models for statistical language modelling[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 641-648.
29. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.
30. Turney P D, Pantel P. From frequency to meaning: Vector space models of semantics[J]. Journal of artificial intelligence research, 2010, 37(1): 141-188.
31. Lebret R, Collobert R. Word Embeddings through Hellinger PCA[J]. EACL 2014, 2014: 482.
32. Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 Association for Computational Linguistics, 2012:873-882.
33. Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 173-180.
34. 李宇明. 语言学概论(第2版)[M]. 高等教育出版社, 2008.
35. 朱德熙. 语法讲义[M]. 商务印书馆, 1982.
36. Charles Carpenter Fries. 1952. The structure of english: An introduction to the construction of English sentences. Language, 31(2):312–345.
37. Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity[C]//AAAI. 2006, 6: 775-780.
38. Niu Z Y, Ji D H, Tan C L. Optimizing feature set for Chinese word sense disambiguation[C]//Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems. 2004.
39. Cai K, Shi X, Chen Y, et al. Chinese Word Sense Induction based on Hierarchical Clustering Algorithm. CLP. 2010.
40. Huang H, Yang Zh, and Jian P. Unsupervised Word Sense Disambiguation Using Neighborhood Knowledge. 25th Pacific Asia Conference on Language, Information and Computation, pages 333–342.
41. Chang C C and Lin C J. 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):1-27.
42. Li W, Lu Q, Li W. Integrating Collocation Features in Chinese Word Sense Disambiguation[C]//Proceedings of the Fourth Sighan Workshop on Chinese Language Processing. pp. 2005: 87-94.

# 致谢

时光荏苒，转眼就要和三年研究生生活挥手告别，三年前第一次来北语参加复试的场景仍旧清晰，恍若昨日。而我也很幸运，在这美好的时光，美丽的校园里，遇见了那么多可爱的人们，经历了很多，也成长了很多，在此对所有关心、帮助和支持我的人表示感谢。

首先衷心感谢我的导师荀恩东教授，感谢荀老师在研究生期间给予的指导与帮助。工程开发上，一直对我严格要求，在开发能力得到提升的同时，也让我对待产品的开发变得精益求精。在学术与科研上，引导的同时，也给予了我充足的空间去探索。我也开始逐渐走进自然语言处理的研究领域，能力不断地得到提升，收获也很多。也感谢荀老师对我转变研究方向以及读博的支持，我才对自己有了新的认识，并从此走上了科研的道路。借此机会向荀老师表示最崇高的敬意和最诚挚的感谢。

感谢实验室的于东老师，有幸与于老师在同一个实验室工作，得到了于老师在学术上更细致的指导，包括实验设计的细节以及论文写作的技巧，每次的讨论都使我收获颇多，在生活上也是备受于老师关照。硕士期间取得的进步跟于老师的指导与鼓励是分不开的。还要感谢实验室的刘颖滨老师、付永刚老师、安维华老师给予的指导与关怀。

感谢实验室的大师兄饶高琦，从入学到现在，一直在学习、科研上给予了我无数的指导和帮助，会为我耐心地解答疑惑，也会在我碰到问题时帮助出谋划策，无愧于“最佳大师兄”的称号。超大的脑洞、超强的“多线程”与“并发”能力、超大的人脉关系网，一直是实验室的师弟师妹们学习的榜样。感谢黄志娥师姐一家以及孟泽豫师姐为研究生生活增添了无数的色彩。感谢梁珊师姐和王昆师兄，与你们一起参与项目很顺利也很愉快。感谢实验室的孙燕南、韩睿方和藏娇娇，与你们一起很开心。感谢实验室的师弟师妹们：杨坤、瞿琴、李琪、徐亦沛、肖晓悦、杨兆勇、吴嘉伟，你们为实验室的生活带来了无限的活力与乐趣。

感谢我的舍友陈志超和李兆杰在我在低谷时的鼓励与引导，以及在生活和学习上的无私地帮助。感谢李飞亚在我身处困境时给予的帮助，还有李琳，高迎明和彭蓓，大家在博士申请和雅思考试时的并肩作战是研究生生活乃至人生中都值得回忆的经历，感谢你们的陪伴。

最后要特别感谢我的家人。感谢我的父母，感谢你们给予我的一切，以及在背后默默地付出。感谢我的姐姐们，你们替我分担着更多的责任。感谢你们在背后默默地支持我，尽可能多地为我创造条件，我才能一步一步地走到现在。

感恩的人与事太多，无法一一言表，谨以此文献给所有关心、帮助和支持我的人。

# 已发表文章目录

1. **唐共波**, 于东, 荀恩东. 基于知网义原词向量表示的无监督词义消歧方法[J]. 中文信息学报, 2015, 6: 24-30.
2. **Tang G**, Guo Y, Yu D, et al. A hybrid re-ranking method for entity recognition and linking in search queries [M]//Natural Language Processing and Chinese Computing. Springer International Publishing, 2015: 598-605.
3. Yu D, Yu C, **Tang G**, et al. An Introduction to BLCU Personal Attributes Extraction System [J]. CLP 2014, 2014: 120-125.
4. 于东, 饶高琦, **唐共波**, 等. 复杂科技术语构词中的语素化[J]. 中国科技术语, 2015, 2: 15-20.

1. http://www.keenage.com/ [↑](#footnote-ref-1)
2. https://www.wikipedia.org/ [↑](#footnote-ref-2)
3. http://ictclas.nlpir.org/ [↑](#footnote-ref-3)
4. http://nlp.stanford.edu/software/tagger.shtml [↑](#footnote-ref-4)
5. https://code.google.com/archive/p/word2vec/ [↑](#footnote-ref-5)
6. http://nlp.stanford.edu/projects/glove/ [↑](#footnote-ref-6)
7. https://dumps.wikimedia.org/ [↑](#footnote-ref-7)
8. http://icl.pku.edu.cn/icl\_groups/corpus/spec.htm [↑](#footnote-ref-8)
9. http://www.cis.upenn.edu/~treebank/ [↑](#footnote-ref-9)
10. https://www.csie.ntu.edu.tw/~cjlin/libsvm/ [↑](#footnote-ref-10)
11. http://wordnet.princeton.edu/ [↑](#footnote-ref-11)
12. https://framenet.icsi.berkeley.edu/fndrupal/ [↑](#footnote-ref-12)
13. http://babelnet.org/ [↑](#footnote-ref-13)