借重于人工知识库的词和义项的向量表示:以 HowNet 为例*

孙茂松[1,2]，陈新雄[1]

（1.清华大学计算机科学与技术系，清华信息科学技术国家实验室，清华大学智能技术与系统国家重点实验室，北京，100084；2. 北京市成像技术高精尖创新中心，首都师范大学，北京 100048）

摘要： 本文旨在以 HowNet 为例，探讨在表示学习模型中引入人工知识库的必要性和有效性。目前词向量多是通过构造神经网络模型，在大规模语料库上无监督训练得到，但这种框架面临两个困难问题：一是低频词的词向量质量难以保证，二是多义词的义项向量无法获得。本文提出了融合 HowNet 和大规模语料库的义原向量学习神经网络模型，并以义原向量为桥梁，自动得到义项向量及完善词向量。初步的实验结果表明该模型能有效提升在词相似度和词义消歧任务上的性能，有助于低频词和多义词的处理。作者指出，借重于人工知识库的神经网络语言模型应该成为今后一段时期自然语言处理的研究重点之一。

关键词：词向量；义项向量；义原向量；HowNet；神经网络语言模型

中图分类号：TP391　　**文献标识码：**A

Embedding for Words and Word Senses based on Human Annotated Knowledge Base: Use HowNet as a Case Study

Maosong Sun[1,2], Xinxiong Chen[1]

(1. State Key Lab. of Intelligent Technology and Systems, National Lab. on Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China)

(2. Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China)

Abstract: This paper aims to address the necessity and effectiveness of encoding a human annotated knowledge base into a neural network language model, using HowNet as a case study. Traditional word embedding is derived from neural network language model trained on a large-scale unlabeled text corpus, however, it suffers from two weaknesses: the first, the quality of resulting vectors of low frequent words is not satisfactory in general, and the second, sense vectors of polysemous words are not available in essence. We propose neural network language models that can systematically learn embedding for all the semantic primitives defined in HowNet, and consequently, obtain word vectors, in particular for low frequent words, and word sense vectors in terms of these semantic primitive vectors. Preliminary experimental results show that our models can improve the performance in tasks of both word similarity and word sense disambiguation. We believe the research for neural network language models incorporating human annotated knowledge bases would be a critical issue deserving our attention in the coming years.

Keywords: word embedding; word sense embedding; sematic primitive embedding; HowNet; neural network language model

---

* **收稿日期：**　　**定稿日期：**

基金项目：国家社会科学基金重大项目（13&ZD190）；国家自然科学基金重点项目（61133012）

作者简介：孙茂松（1962 年——），男，教授，计算语言学，机器学习，互联网智能；陈新雄（1988 年——），男，博士，自然语言处理。

1 引言

词向量表示旨在学习词的低维实数向量表示，是自然语言处理的重要任务之一。训练得到的词向量可直接用于计算两个词之间的语义相关性，同时可作为特征广泛应用于诸多后续的自然语言处理任务中，如信息检索、语言模型、词义消歧、词义组合和命名实体识别等。

目前的词向量一般都是在极大规模生语料库（对中文需要经过基本的分词处理）上通过构建神经网络语言模型以无监督学习的方式训练得到。这种典型的计算框架存在两个“天然”的缺陷，或者说困难问题。第一个困难问题是：经验表明，低频词的词向量的语义表达质量较高频词会显著下降，很多情况下难以令人满意；第二个困难问题是：词汇中很多词是多义的，但从生语料库中根本不可能学习到多义词的义项向量表示，其在词义消歧、词义组合等后续任务中的效用会大打折扣。

显然，不借助于其它资源是无法解决上述两个天然“缺陷”的。关于第一个困难问题：根据齐夫定律，必然存在一个数量十分庞大的低频词集合，所以无论语料库规模多大，这些词的词向量的语义表达质量问题始终会如“梦魇”相随；关于第二个困难问题：如果有一个经过人工义项标注的极大规模语料库，词的义项向量表示问题在典型框架下将会迎刃而解，但人工标注这样一个语料库投入巨大，并不现实（进一步地，即使有了这样一个语料库，低频义项的向量质量问题还是会无可避免地凸显出来）。在词和义项的向量表示学习中系统性地借重于其它资源，尤其是人工业已建立起来的大规模知识库，无疑是我们攻坚克难的一条可行之道。而现实存在着的若干高质量的人工标注知识库（英文如 WordNet，中文如 HowNet[1]等）中，蕴含了十分丰富的关于语言和世界的知识（实际上体现了一流专家从认知或计算角度对语言和世界的系统化认识），如何将这些知识有效合理地加入到词向量和义项向量学习中，便成为了表示学习中的一个重要课题。

已有一些研究者将人工知识库与词向量或义项向量的学习进行了结合。如：Wang 等[2]提出利用机器学习中的正则化（regularization）技术将词汇的语义关联度作为正则化因子嵌入到词向量学习的优化目标中，使得学到的词向量融合了先验知识（例如两个词是同义词）。Chen 等[3]利用 WordNet 为多义词的不同义项训练相应的义项向量，有效提升了英文词义消歧的效果；Rothe 等[4]利用 WordNet 将词向量自动扩展到 Synset 向量上；唐共波等[5]基于 HowNet 中的基本语义单位——“义原”——来学习义项的向量表示，用于无监督词义消歧。

本文以 HowNet 为例，研究如何将人工知识库的信息加入到词向量和义项向量的学习过程中。我们提出了 HowNet 和极大规模生语料库共同作用的义原向量学习方法，并以学到的义原向量为桥梁，求出义项向量，完善化词向量，以期对解决前文提及的两个困难问题都有所裨益。尽管唐共波等已经提出了为 HowNet 义原学习向量的思路，但其策略较为简单：根据 HowNet 全部词（超过 10 万个）中的 35,247 个单义原词（约占全部词的 33.75%），将北京语言大学中文语料库 BCC（规模为 13 亿字左右）中的单义原词全部替换为义原，得到 182,398 个义原实例，然后利用经典的 word2vec 在替换处理后的 BCC 上同时构造义原向量和词向量。需要注意的是，义原实例仅占了 BBC 极小一部分，这提示义原向量的训练可能很不充分，并且该方法能得到向量表示的义原也只占全部义原的 60.95%，因而通过对这些义原向量求平均之类操作而得到的义项向量应该是相对粗放的（也不能保证 HowNet 中的每一个义项都能得到向量表示）。而我们提出的方法囊括了 HowNet 中的全部词和全部义原，设计了更为复杂、周详的模型。

本文安排如下：第二节简要介绍 HowNet 中词的形式化描述系统以及我们构造的两类基于 HowNet 的义原向量表示学习神经网络模型，即义项不敏感的模型和义项不敏感的模型；第三节针对词相似度任务和词义消歧任务验证我们所提模型的有效性；第四节从最近邻视角对实验结果进行了具体观察，第五节归纳并强调了我们的基本观点。

2 借重于 HowNet 的词和义项的向量学习模型

2．1 HowNet 中词的形式化描述系统

HowNet 是使用最为广泛的可计算中文语义词典。在 HowNet 中，词的形式化描述系统是按照词-义项-义原三层结构来组织的。即词按照义项分列，义项又被作者精心设计的义原所定义（义原可以理解成功能类似化学中“元素”的中文基本语义单位，所有义项均由义原的不同组合而成）。表 1 给出了“包袱”一词在 HowNet 中的形式化描述。

表 1 “包袱”一词在 HowNet 的形式化描述

No.=015240 W_C=包袱 G_C=noun E_C=不要背～，思 想～，在心里卸下～，扔下心里的～ W_E=burden G_E=noun E_E= DEF=duty|责任	No.=015243 W_C=包袱 G_C=noun E_C=两个～,背上～，打开～，捆好～ W_E=parcel G_E=noun E_E= DEF={physical|物质：{wrap|包扎}}	No.=015245 W_C=包袱 G_C=noun E_C=用～包起来，用～裹 W_E=cloth-wrapper G_E=noun E_E= DEF={tool|用具：{wrap|包扎}}

其中 No.是每一个义项在 HowNet 中的序号，W_C、G_C、E_C、W_E、G_E、E_E 分别表示中文词语、中文词性、中文示例、英文词语、英文词性、英文示例，DEF 定义了相应的义项，如这里“包袱”一词的第一个义项 No.=015240（表示 “心理负担”的意思）是使用义原“duty|责任”来描述的。

2．2 义项不敏感的义原向量表示学习神经网络模型

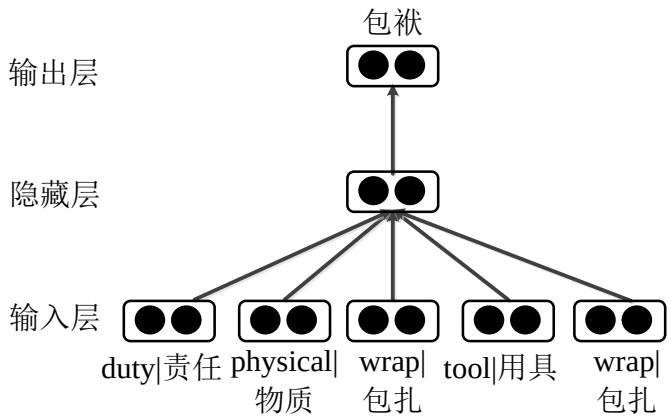

图 1 义项不敏感的义原向量表示学习神经网络模型

我们构建了一个如图 1 所示的神经网络来学习义原向量。这个模型与 Mikolov 的 CBOW 模型貌似差不多，但实际上还是有相当差别的：CBOW 模型同步更新上下文向量与词向量，而我们的模型首先在训练语料库上运行 CBOW 后得到词的向量表示，然后固定训练好的词向量不变，不断更新所辖的义原向量。其基本思想是：训练词所辖的义原向量去逼近该词向量（注意：此时各义项的义原被不加区别地独立排列，故称之义项不敏感），使学到的义原向

量可较好的预测这些义原共同作用所定义的词向量。

形式化地，给定词w_i的词向量$\vec{w}_i$和该词对应的义原向量$\vec{p}_1,\ldots,\vec{p}_m$，训练目标为：

$$L = \frac{1}{T}\sum_{i=1}^{T}\log \Pr(w_i \mid p_1,\ldots,p_m) \tag{1}$$

求和遍历了整个训练集（规模为 T）来计算义原正确预测所定义词的对数概率。

我们使用 softmax 函数来定义预测的概率$\Pr(w_i \mid p_1,\ldots,p_m)$：

$$\Pr(w_i \mid p_1,\ldots,p_m) = \frac{\exp(\vec{p}^{\mathrm{T}}\cdot\vec{w}_i)}{\sum_{w'\in W}\exp(\vec{p}^{\mathrm{T}}\cdot\vec{w}')} \tag{2}$$

其中W是词表，$\vec{w}_i$是目标词w_i的向量表示，$\vec{p}$是所有义原向量的平均值：

$$\vec{p} = \frac{1}{m}\sum_{j=1}^{m}\vec{p}_j \tag{3}$$

以“包袱”为例，我们的模型会把它的所有义原（即“duty|责任”，“physical|物质”，“wrap|包扎”，“tool|用具”和“wrap|包扎”）的向量的平均作为隐藏层的向量，用于预测“包袱”一词。

从公式（2）可以看到，计算预测概率时需要遍历整个词表，而词表大小往往是比较大的（这里超过 10 万个词），因此本文使用了层次化的 softmax 去降低计算复杂度。

对于词的迭代训练有两种不同的选择：

（1）在 HowNet 词典上进行迭代（遍历的训练集为 HowNet 词典）。即遍历词典中的每一个词，所有词的训练次数都一样(此时义原向量的更新过程与语料库无关)。

（2）在大规模语料库上进行迭代（遍历的训练集为语料库）。即依次遍历大规模语料库中的每一个词,在一轮训练过程中,每一个词的训练次数就是这个词在语料库中出现的次数。

2．3 义项敏感的义原向量表示学习神经网络模型

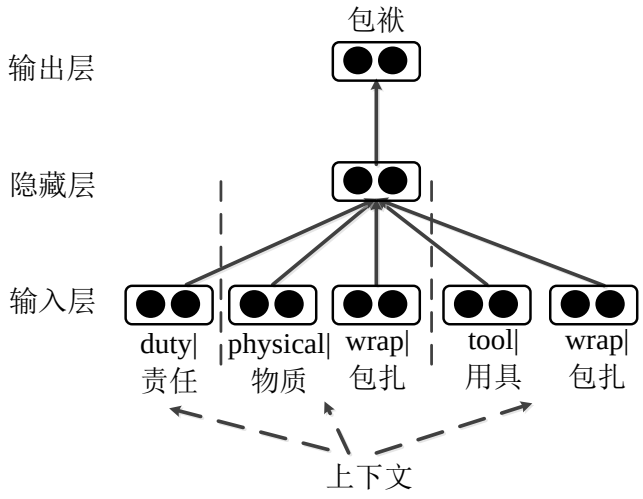

图 2 义项敏感的义原向量表示学习神经网络模型

义项不敏感的义原向量表示学习神经网络模型直接使用所有义原向量来预测所定义的词，在更新义原向量时也会更新所有的义原，这个策略显得有些“简单粗暴”，因为一个词在句子的具体上下文中会取不同的义项，也就是说，不会是所有的义项同时在起作用。于是我们进一步提出了一种义项敏感的义原向量表示学习神经网络模型（图 2），使得在学习过程中模型会根据句子的具体上下文来选择w_i最可能的义项，然后只使用相对应的义原向量来预测w_i，同样地，也只更新相应的义原向量（注意，这个训练过程是动态的；在训练过程中基于 CBOW 预处理得到的词向量依然是始终固定不变）。

形式化地，给定词w_i的词向量$\vec{w}_i$和对应的义原向量$\vec{p}_1,\ldots,\vec{p}_m$，首先计算词$w_i$的第$j$个义项的向量$\vec{w}_{ij}$：

$$\vec{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\vec{p}_k \tag{4}$$

然后挑选与该词的上下文最相近的义项：

$$\vec{w}_{ir} = \arg\max\cos(\vec{w}_{ij},\vec{c}) \tag{5}$$

其中$\vec{c}$是该词在当前上下文窗口中词向量的平均。

当挑选出与上下文最相近的义项之后，在预测（公式 3）和更新时都只使用这个义项所对应的若干义原，而不是该词的所有义原。

同样以“包袱”为例，在训练时，模型会根据该词在语料库中的当前上下文来选择最为相关的义项，不妨假设某个时刻选择了图 2 中的第二个义项，那么模型将使用义原“physical|物质”和“wrap|包扎”的平均向量来预测“包袱”，对应地，也只会更新义原“physical|物质”和“wrap|包扎”的向量。

2．4 义项向量和词向量的获得

经过 2.2 或 2.3 的处理后，我们得到了所有义原的向量表示。则：（1）各义项向量取所辖各相应义原向量的平均即可。（2）对较高频词，其词向量即取 CBOW 预处理得到的词向量，而对较低频词，我们认为 CBOW 预处理得到的词向量的可信度值得商榷，于是乎舍之而取其所辖义原向量的平均作为词向量（绝大多数情况下都是单义项词）。

3 实验

本节针对两个任务来验证我们所提出的模型：一个是词相似度任务以检验得到的词向量的有效性，另一个是标准的词义消歧任务以验证基于义项向量的消歧算法。实验结果表明本文提出的模型：(1)在词相似度任务上能够提升与人类打分的相关性；(2)在一个标准词义消歧任务中能超过现有的最好无监督消歧系统。

实验使用搜狗实验室提供的 SogouT 互联网语料库[1]作为训练语料库。SogouT 共包含来自互联网各种类型的 1.3 亿个原始网页，大小超过 5TB。首先预处理去掉网页内的噪声内容，如标签、链接、脚本等，得到纯中文网页正文文本 152.8GB，超过 19 亿个句子，共 554 亿

[1] http://www.sogou.com/labs/dl/t.html

个字符，其中汉字(不含标点)超过 478 亿个。句子去重后得到 7 亿个不同的句子，共 256 亿个字符，其中汉字(不含标点)221 亿个，共 72GB。然后使用 THULAC[2]对语料库进行自动分词和词性标注。THULAC(THU Lexical Analyzer for Chinese)是由清华大学自然语言处理与社会人文计算实验室研制的一套中文词法分析工具包，对开放文本具有很强的分词和词性标注功能，可自由下载。

我们使用 HowNet 2012 版本[3]作为义项词典，经整理后，共含 103,843 个词，128,578 个义项，2,157 个义原。词和义原的向量维度均设置为 200（义项向量的维度因此也是 200）。

3．1 词相似度任务

我们采用公开数据集 wordsim240 测试词向量的质量。共有 240 个词对，每一个词对都赋以 10 个人工相似度打分(打分范围为 0-10)。

实验中两个词(u，v)之间的相似度计算如下：

$$AvgSim(u,v) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} d(\vec{u}_i, \vec{v}_j) \qquad (6)$$

$d(\vec{u},\vec{v})$是计算两个向量的相似度函数，这里使用余弦相似度。

表 2 给出了各模型得到的词间相似度与人工打分的 Spearman 相关系数。其中 CBOW，Skip-gram 和 GloVe 模型作为比较的基线。CBOW 和 Skip-gram 模型是 Mikolov 等的神经网络模型，GloVe 是 Pennington 等[7]的矩阵分解模型，都是词向量表示中的经典模型。

表 2 不同模型在数据集 wordsim240 词相似度任务上的评测结果

模型	与人工打分的相关系数*100
CBOW	55.85
Skip-gram	53.42
GloVe	48.22
义项不敏感（遍历的训练集为 HowNet 词典）	56.93
义项不敏感（遍历的训练集为语料库）	57.48
义项敏感（遍历的训练集为语料库）	57.03

实验结果初步显示：

（1）即使在“义项不敏感（遍历的训练集为 HowNet 词典）”的配置下，我们的模型效果也能比所有的基线模型（CBOW，Skip-gram 和 GloVe）要好。分析其原因，我们发现：通过义原向量来预测词向量的做法对于较高频词并没有明显提升，因为这些词在基线模型中已经得到了非常充分的训练，但是对于较低频词，我们的模型能够通过对应义原在较高频词中的训练来提升较低频词的向量质量，从而达至更好的效果。

（2）在“义项不敏感（遍历的训练集为语料库）”的配置下，大规模语料库上的训练使词的更新次数正比于其出现的频度，这导致高频词对应的义原得到更充分的训练，因此实验效果得以进一步的提升。

（3）在“义项敏感（遍历的训练集为语料库）”的配置下，这种理论上更“精致”的模型并未如愿取得比（2）更好的实验效果。

---

[2] http://thulac.thunlp.org/

[3] http://www.keenage.com/html/c_index.html

3．2 词义消歧任务

我们从 SemEval2007 中文词义消歧任务的公开数据集中选取了 6 个词(“把握”,“材料”,“老”,“没有”,“突出”和“研究”)的 96 个实例作为测试集，以测试义项向量的表现。

我们采用了随机选择义项、Li 等[8]的朴素 Bayes 分类，Wang 等[9]的 PageRank（目前中文词义消歧任务上最好的无监督学习方法之一。该方法首先根据 HowNet 中义原的树状结构来构建一个图，然后在这个图上运行 PageRank 算法得到最终的消歧结果）作为基线模型。表 3 给出了不同模型在数据集上取得的实验结果。

表 3 不同模型在 SemEval2007 中文词义消歧任务上的评测结果

模型	平均准确率
随机选择义项	0.24
Li 等的朴素 Bayes	0.44
Wang 等的 PageRank	0.54
义项不敏感(遍历的训练集为 HowNet 词典)	0.56
义项不敏感（遍历的训练集为语料库）	0.57
义项敏感（遍历的训练集为语料库）	0.58

实验结果初步显示：

（1）我们提出的各个模型都都比目前最好的 PageRank 模型效果要好。

（2）与词相似任务时的情况略有不同，“义项敏感（遍历的训练集为语料库)”配置取得了比“义项不敏感（遍历的训练集为语料库)”稍好一点的实验效果。

4 最近邻视角下的实例观察

这里基于最近邻视角，通过观察若干实例加深对实验结果的感性认识。实际上，由于我们的模型生成的义原向量、义项向量和词向量都是在一个空间的，所以义原、义项和词这三者之间，借助义原这个桥梁是完全打通的，在语义计算上可以自由“穿越”。

“穿越”可以沿着任意一个方向“由此及彼”。如表 4 显示了与给定义项和义原最近邻的词的实例（通过计算相应向量之间的夹角余弦)。可以看出所生成的义项向量和义原向量具有一定的合理性（呼应本文开篇所说的第二个困难问题)。

表 4 义项向量和义原向量的最近邻词示例

义项或义原	最近邻词
包袱(义项 1)	责任，责无旁贷，义不容辞，重责，守土有责
duty|责任	责任，责无旁贷，义不容辞，重责，守土有责
包袱(义项 2)	纸卷，装袋，纸箱，包装，油纸
physical|物质	铁磁，电导，电导率，基态，表征
wrap|包扎	捆扎，塑料纸，布条，包装纸，抖开
包袱(义项 3)	抖开，红绸子，绸布，捆扎，油布
tool|用具	光闪闪，红绸子，抖开，放置，鼓弄
wrap|包扎	捆扎，塑料纸，布条，包装纸，抖开

“穿越”也可以“由己及己”。如表 5 显示了与给定义原最近邻的义原。

表 5 义原向量之间最近邻示例

义原	最近邻的义原
duty|责任	bear|承担，effortful|费力，GoodSocial|好风气，affairs|事务，trusty|可信
physical|物质	artifact|人工物，entity|实体，thing|万物，animate|生物，inanimate|无生命
tool|用具	implement|器具，shape|物形，fittings|配件，decorate|装饰，mark|标志
wrap|包扎	fold|摺叠，twine|打结，weave|辫编，bend|折弯，straighten|拉直

最后我们还是回到词向量，进行一次最为期待的“由己及己”的“穿越”：给定一个词，尤其是较低频词，观察其最近邻词（呼应本文开篇所说的第一个困难问题）。下面我们进一步来看使用这些义项向量来模拟低频词的向量的结果。

表 6 低频词的最近邻词示例

词	SogouT 词频	CBOW 给出的最近邻词	我们的模型给出的最近邻词
背债	99	替前夫、家欠、堕胎费	债款、债务、债项
蠢笨	95	自大无比、愚蠢无知、懦弱无用	木讷、呆头呆脑、愚蠢
二赖子	51	横眉瞪目、张五魁、花荣志	恶棍、穷凶极恶、逞凶
匡谬	10	吉金录、曲话、笋谱	错误、订正、讹误
不宣而战	1	西方、伊朗人、冲突国	杀敌、拔寨、整军

表 6 显示，经典的 CBOW 模型对于这些词的训练效果并不好，而我们的模型通过义原向量却可以有效捕捉到低频词的语义（其中“匡谬”一例最为典型）。我们注意到，HowNet 全部 103,843 个词中，在 SogouT 中频度小于 100 的词有 35,274 个（超过 33%），可见我们的模型的受益面是相当大的。

5 结束语

本文的主要目的是以 HowNet 为例，探讨并强调在表示学习模型中引入人工知识库的必要性和有效性。“几乎从零开始”(almost from scratch)是神经网络语言模型所标榜、推崇的一种学习方式，也是其大“秀”自己强大学习能力“肌肉”的一种展示方式。但是必须清醒地认识到，这种方式并不能包打天下，对某些类型的任务不是“自足”的（如本文的义项向量学习任务），也存在其“阿喀琉斯之踵”（如本文的低频词向量学习任务），再强大的力量也无法自己举起自己，必须借助“外力”才能摆脱其局限性。而各类人工知识库就是我们必须依赖同时也是可以依赖的“外力”。一个充分融合了人工知识库(理想状态应该是统筹了多个相关人工知识库，包括语言知识库和世界知识库)的神经网络语言模型能以一种无监督学习的方式坐收基本面上“几乎从零开始”和特定任务上“站在巨人的肩膀上”之利，往往能够避免或者大大缓解新的人工投入，从而取得事半功倍之效。

科学合理地构造此类模型不是轻而易举的，需要匠心独运，如 HowNet 中义原系统的结构信息在本文提出的模型中就还没有用上。此类工作应该成为我们今后研究的重点。

References

.

[1]. 董强, 董振东,《知网》, in http://www.keenage.com.

[2]. Wang, Y., Z. Liu and M. Sun, Incorporating Linguistic Knowledge for Learning Distributed Word Representations. PloS one, 2015. 10(4): p. e0118437.

[3]. Chen, X., Z. Liu and M. Sun, A Unified Model for Word Sense Representation and Disambiguation, in Proceedings of EMNLP. 2014. p. 1025--1035.

[4]. Rothe, S.A.S.U., AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015. p. 1793--1803.

[5]. 唐共波, 于东与荀恩东, 基于知网义原词向量表示的无监督词义消歧方法. 中文信息学报, 2015. 29(6): 第23-29页.

[6]. Mikolov, T., W. Yih and G. Zweig, Linguistic Regularities in Continuous Space Word Representations, in HLT-NAACL. 2013. p. 746--751.

[7]. Pennington, J., R. Socher and C.D. Manning, Glove: Global vectors for word representation, in Proceedings of EMNLP. 2014.

[8]. Li, W. and A. McCallum, Semi-supervised Sequence Modeling with Syntactic Topic Models, in Proceedings of AAAI. 2005. p. 813.

[9]. Wang, J., J. Liu and P. Zhang, Chinese Word Sense Disambiguation with PageRank and HowNet, in Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

孙茂松（1962年——），男，教授，计算语言学，机器学习，互联网智能。

E-mail: sms@mail.tsinghua.edu.cn

陈新雄（1988年——），男，博士，自然语言处理

E-mail: amiucxx@gmail.com