从词到词义嵌入：语义向量表征综述

Jose Camacho-Collados camachocolladosj@cardiff.ac.uk

计算机科学与信息学院 卡迪夫大学 英国

Mohammad Taher Pilehvar pilehvar@iust.ac.ir

计算机工程学院 伊朗科技大学 伊朗德黑兰

摘要

在过去的几年中，分布式语义表征已被证明是有效且灵活的先验知识保持者，可以集成到下游应用程序中。本篇综述侧重于语义表征。我们从词向量空间模型背后的理论背景开始，并强调它们的主要限制之一：含义混淆缺陷，这是由于将具有所有可能含义的词表征为单个向量而引起的。然后，我们解释了如何通过从词级别过渡到更细粒度级别的词义（在其更广泛的接受范围内）作为建模明确词汇含义的方法来解决这一缺陷。我们对语义表征的两个主要分支（即无监督和基于知识）中的广泛技术进行了全面概述。最后，本篇综述涵盖了此类表征的主要评估程序和应用，并对其四个重要方面进行了分析：可解释性、感觉粒度、对不同领域的适应性和组合性。

1. 引言

近来，处理大量文本数据以将词的语义嵌入到低维向量中的基于神经网络的方法，即所谓的词嵌入，引起了很多关注（Mikolov、Chen、Corrado & Dean，2013a；Pennington , Socher & Manning, 2014)。词嵌入已经证明了它们在存储有价值的句法和语义信息方面的有效性（Mikolov, Yih, & Zweig, 2013d）。事实上，它们已被证明胜任多项自然语言处理 (NLP) 任务，主要是由于它们的泛化能力 (Goldberg, 2016)。广泛的应用展示了集成词嵌入的改进，包括机器翻译 (Zou, Socher, Cer, & Manning, 2013)、句法解析 (Weiss, Alberti, Collins, & Petrov, 2015)，文本分类 (Kim, 2014) 和问答 (Bordes, Chopra, & Weston, 2014)，仅举几例。

然而，尽管它们在捕获单词的语义属性方面具有灵活性并获得成功，但词嵌入的有效性通常受到一个重要限制的阻碍，我们将其称为含义混淆缺陷：无法区分单词的不同含义。一个词可以有一种含义（单义）或多种含义（歧义）。例如，名词nail[1]可以根据上下文指代两种不同的含义：手指的一部分或金属物体。因此，名词 nail 被称为是模棱两可的。一个模棱两可的词的每一个单独的意义被称为词义，而列出词的不同意义的词汇资源通常被称为语义清单。[2] 虽然一般语义清单（例如 WordNet）中的大多数词通常是单义的[3]，根据词的经济通用性原则（Zipf，1949），频繁出现的词往往具有更多的意义。因此，准确捕捉歧义词的语义对于 NLP 系统的语言理解起着至关重要的作用。

为了解决含义混淆缺陷，许多方法尝试对单个词义进行建模。在本篇综述中，我们试图综合与语义表征学习最相关的工作。这些方法的主要区别在于它们如何对意义进行建模以及从何处获得意义。无监督模型直接从文本语料库中学习词义，而基于知识的技术利用词汇资源的语义清单作为其表征意义的主要来源。在本篇综述中，我们涵盖了这两类用于学习分布式语义表征的技术，包括评估程序和对其主要属性的分析。虽然该调查旨在尽可能广泛，但鉴于所审查主题的广度，某些领域可能没有得到足够的覆盖面，无法完全独立。但是，对于这些案例，我们为有兴趣了解该主题的读者提供了相关的指导。鉴于本篇综述旨在覆盖广泛的受众，我们已努力使其尽可能易于理解。因此，技术细节可能不一定提供完整的细节，而是它们背后的直觉。

本篇综述的其余部分结构如下。首先，在第 2 节中，我们提供了词义的理论背景、它们是什么、为什么对它们进行建模可能有用以及它的主要范式。然后，在第 3 节中，我们描述了直接从文本语料库中学习的无监督意义向量建模技术，而在第 4 节中，解释了与词汇资源相关的表征。第 5 节介绍了常见的评估程序和基准，第 6 节介绍了在下游任务中的应用。最后，我们在第 7 节介绍了无监督和基于知识的表征之间的分析和比较，第 8 节介绍了主要结论和未来挑战。

2. 背景

本节提供了支持从单词级别向更细粒度级别的词义和概念转变的理论基础。首先，我们论述向量空间模型的背景，特别是词表征学习（第 2.1 节）。然后，我们解释了导致语义建模技术发展的单词表征的一些主要缺陷（第 2.2 节），并描述了表征语义的主要范式（第 2.3 节）。在 2.4 节中，我们简要介绍了词义消歧相关任务的历史背景。最后，我们解释了本篇综述所遵循的符号（第 2.5 节）。

2.1 词表征学习

自 NLP 出现以来，词表征学习一直是语义学的主要研究领域之一。我们首先介绍基于向量空间模型的词表征学习背后的主要理论（第 2.1.1 节），然后转向学习词嵌入的新兴理论（第 2.1.2 节）。

2.1.1 向量空间模型

单词表征学习最突出的方法之一是基于向量空间模型 (VSM)，它得到了人类认知研究的支持 (Landauer & Dumais, 1997; Gärdenfors, 2004)。最早应用于 NLP 的 VSM 将文档视为一个向量，其维度是整个词汇表 (Salton, Wong, & Yang, 1975)。根据文档中的词频初始化计算各个维度的权重。在后续的研究中，探索了不同的权重计算指标，但主要基于频率或归一化频率（Salton & McGill，1983）。该方法已成功地改进并应用于各种 NLP 应用程序，例如信息检索 (Lee, Chuang, & Seamons, 1997)、文本分类 (Soucy & Mineau, 2005) 或情感分析 (Turney, 2002) 等等。Turney 和 Pantel (2010) 全面概述了 VSM 及其应用。

基于文档的 VSM 也已扩展到其他词汇项，如单词。在这种情况下，一个词通常表示为向量空间中的一个点。传统上，基于语料库中共现词的归一化频率构建了基于词的向量 (Lund & Burgess, 1996)，遵循 Harris (1954) 的初始理论。词 VSM 背后的主要思想是共享相似上下文的词应该在向量空间中接近（因此，具有相似的语义）。图 1 显示了一个单词 VSM 的示例，其中清楚地突出了这个潜在的邻近公理。

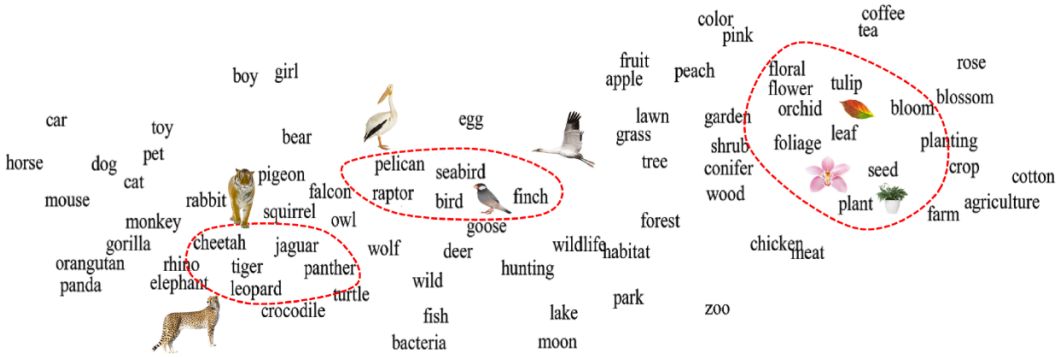


图 1：使用 t-SNE 将样本词向量空间的子集缩减为二维（Maaten & Hinton，2008）。在语义空间中，具有相似含义的单词往往出现在彼此接近的位置，如与大型猫科动物、鸟类和植物相关的这些单词簇（由红色虚线分隔）所突出显示的那样。

基于向量的表征已经确立了它们在 NLP 任务中的有效性，例如信息提取 (Laender, Ribeiro-Neto, da Silva, & Teixeira, 2002)、语义角色标记 (Erk, 2007)、词相似性 (Radinsky, Agichtein, Gabrilovich, & Markovitch, 2011)、词义消歧 (Navigli, 2009) 或拼写校正 (Jones & Martin, 1997) 等。传统 VSM 方法的主要缺点之一是生成的向量的高维数。由于维度对应于词汇表中的单词，因此这个数字很容易达到数十万甚至数百万，具体取决于底层语料库。一种常见的降维方法使用奇异值分解 (SVD)，也称为潜在语义分析 (Hofmann, 2001; Landauer & Dooley, 2002, LSA)。此外，最近的模型还利用神经网络直接学习低维单词表征。这些模型将在下一节中介绍。

2.1.2 词嵌入

从文本语料库中学习低维向量也可以通过利用神经网络来实现。这些模型通常被称为词嵌入，并且由于它们的泛化能力已被证明可以提供有价值的先验知识（Goldberg，2016）。事实证明，当集成到神经网络架构中时，该属性对于在许多 NLP 任务中实现最先进的性能至关重要（Zou et al., 2013；Kim，2014；Bordes et al., 2014；Weiss et al., 2015)。

这个较新的预测分支，其架构基于优化某个目标（Bengio、Ducharme、Vincent & Janvin，2003；Collobert & Weston，2008；Turian、Ratinov & Bengio，2010；Collobert、Weston、Bottou、Karlen、Kavukcuoglu , & Kuksa, 2011），通过 Word2vec (Mikolov et al., 2013a) 得到普及。Word2vec 基于一个简单但高效的架构，提供有趣的语义属性（Mikolov et al., 2013d）。提出了两种不同但相关的 Word2vec 模型：Continuous Bag-OfWords (CBOW) 和 Sikp-gram。 CBOW 架构基于前馈神经网络语言模型 (Bengio et al., 2003)，旨在使用其周围上下文预测当前单词，最小化以下损失函数：

$$E = -\log\left(p\left(\vec{w}_t \mid \vec{W}_t\right)\right) \qquad (1)$$

其中w_t是目标词，$W_t = w_{t-n}, \ldots, w_t, \ldots, w_{t+n}$表示上下文中的词序列。图 2 显示了 Word2vec 的 CBOW 和 Skip-gram 模型的一般架构的简化。该架构由输入层、隐藏层和输出层组成。输入层具有词表的大小，并将上下文编码为给定目标词的周围词的单热向量表示的组合。输出层与输入层大小相同，在训练阶段包含目标词的 one-hot 向量。 Skip-gram 模型类似于 CBOW 模型，但在这种情况下，目标是在给定目标词的情况下预测周围上下文中的词，而不是预测目标词本身。有趣的是，Levy 和 Goldberg (2014b) 证明了 Skip-gram 实际上可以被视为点互信息 (PMI) 共现矩阵的隐式分解。

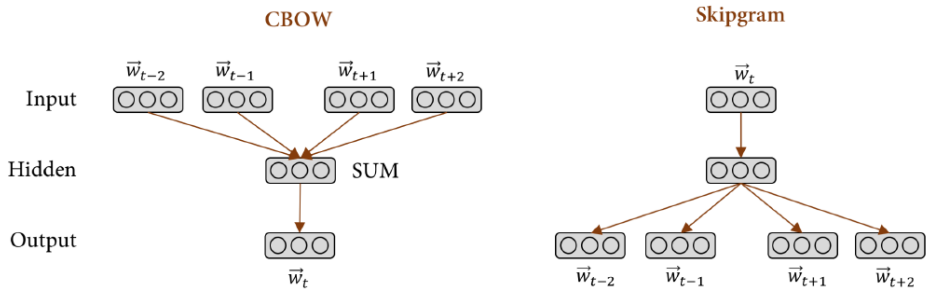


图 2：Word2vec 的 CBOW 和 Skipgram 模型的学习架构（Mikolov 等人，2013a）

另一个突出的词嵌入架构是 GloVe (Pennington et al., 2014)，它通过双线性回归模型结合了全局矩阵分解和局部上下文窗口方法。近年来，已经提出了更复杂的方法来尝试提高词嵌入的质量，包括利用依赖分析树 (Levy & Goldberg, 2014a) 或对称模式 (Schwartz, Reichart, & Rappoport, 2015) 的模型，利用子词单元(Wieting, Bansal, Gimpel, & Livescu, 2016; Bojanowski, Grave, Joulin, & Mikolov, 2017)，将单词表示为概率分布 (Vilnis & McCallum, 2015; Athiwaratkun & Wilson, 2017; Athiwaratkun, Wilson, & Anandkumar, 2018)、学习多语言向量空间中的词嵌入（Conneau、Lample、Ranzato、Denoyer & Jégou，2018；Artetxe、Labaka & Agirre，2018），或利用知识资源（有关此类型的更多详细信息，请参见第 4.2 节）。[4]

2.2 意义合并缺陷

将每个单词类型表示为语义空间中的一个点的主要目标有一个主要限制：它忽略了单词可以具有多种含义并将所有这些含义混为一个表示的事实。 Schütze (1998) 的工作是最早发现词向量含义混淆缺陷的工作之一。将不同的（可能不相关的）含义合并为单个单独的表示可能会妨碍对以这些为核心的 NLP 系统的语义理解。事实上，词嵌入已被证明无法有效捕捉单词的不同含义，即使这些含义出现在基础训练语料库中（Yaghoobzadeh & Schütze，2016）。含义合并可能对准确的语义建模产生额外的负面影响，例如，在语义空间中，与单词的不同含义相似的语义不相关的单词被拉向彼此 (Neelakantan, Shankar, Passos, & McCallum, 2014; Pilehvar & Collier，2016）。举例来说，两个语义无关的词 *rat* 和 *screen* 在语义空间中被拉向彼此，因为它们与*mouse*的两种不同词义相似，即啮齿动物和计算机输入设备。参见图 3 的说明。[5] 此外，合并缺陷违反了欧几里得空间的三角不等式，这会降低词空间模型的有效性（Tversky & Gati，1982）。为了缓解这一缺陷，过去几年出现了一个新的研究方向，它试图直接对单词的个体含义进行建模。在本篇综述中，我们关注这个新的研究分支，它在词表征学习方面有一些相似之处和特点。

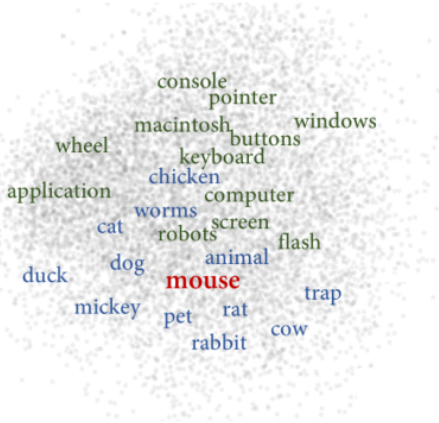


图 3：在模棱两可的词 *mouse* 周围的 2D 语义空间中的含义混淆缺陷的图示。将具有不同含义的单词表示为单个点（向量）会导致将语义上不相关的单词组合在一起，例如*computer*和*rabbit*。

2.3 语义表征

解决词嵌入的含义混淆缺陷的一种解决方案是将单词的单个含义（即词义）表示为独立的表征。这样的表示通常被称语义表征。语义表征技术可以根据词义区分的方式进行广泛分类。将单词的含义划分为多种含义的最佳方式一直是争论的焦点（Erk, McCarthy, & Gaylord, 2009; Erk, 2012; McCarthy, Apidianaki, & Erk, 2016）。传统上，与词义消歧（见第 2.4 节）一样，计算技术依赖于人类产生的固定语义清单，例如 WordNet (Miller, 1995)。语义清单[6]是一种词汇资源，例如字典或同义词库，它为每个单词列出了它可能具有的含义。词义区分也可以通过词义归纳来定义，即通过分析单词出现的上下文自动识别单词的意义。

一般来说，根据词义区分的定义方式，语义表征可以分为两种主要范式：

- **无监督** 在这些表征模型中，语义差异是通过分析文本语料库来诱导的。这种范式与词义归纳非常相关。
- **以知识为基础** 这些技术表征由外部语义清单定义的词义，例如 WordNet.[7] 在这种情况下，最直接相关的任务是词义消歧

在第 3 节和第 4 节中，我们将提供每种范式及其变体的详细信息。

2.4 词义消歧

词义消歧（WSD）是一项与词义混淆缺陷密切相关的任务。 WSD 一直是 NLP 和 AI 中的一项长期任务（Navigli，2009），可以追溯到 20 世纪上半叶，当时它被视为机器翻译的关键中间任务（Weaver，1955）。给定上下文中的单词，WSD 的任务包括将单词与其最合适的词义相关联，该词义由语义清单定义。例如，在句子“My *mouse* was broken, so I brought a new one yeaterday.”中，*mouse*将与它的计算机设备词义相关联，假设在预定义的语义清单中存在这种词义的条目。

WSD 已被归类为 AI 完全问题（Mallery，1988；Navigli，2009），其挑战（如今仍然存在）是多方面的：举几个例子，意义粒度、语料库域或词义的表示（本篇综述中涉及的主题）。此外，WSD 依赖知识资源这一事实带来了额外的挑战，例如此类资源的创建和词义注释语料库的构建。所有这些都代表了一项非常昂贵且耗时的工作，需要针对不同的资源和语言分别完成，并随着时间的推移进行更新。这导致了所谓的知识获取瓶颈（Gale, Church, & Yarowsky, 1992）。

传统上用于 WSD 的知识资源和语义清单与标准词典中的条目相关联，WordNet (Miller, Leacock, Tengi, & Bunker, 1993) 是 WSD 的事实上的语义清单。然而，在实践中可以（并且正在）考虑其他机器可读结构。例如，不断更新的维基百科可以被视为一种语义清单，其中每个条目对应于不同的概念或实体（Mihalcea & Csomai，2007）。甚至可以使用无监督的方法从语料库中自动诱导词义，这项任务称为词义诱导或辨别。

执行 WSD 的方法大致可分为两类：监督（Zhong & Ng, 2010; Iacobacci, Pilehvar, & Navigli, 2016; Yuan, Richardson, Doherty, Evans, & Altendorf, 2016; Raganato, Delli Bovi, & Navigli, 2017b；Luo、Liu、Xia、Chang 和 Sui，2018）和基于知识的（Lesk，1986；Banerjee 和 Pedersen，2002；Agirre、de Lacalle 和 Soroa，2014；Moro、Raganato 和 Navigli，2014；Tripodi & Pelillo，2017；Chaplot & Salakhutdinov，2018 ）。监督方法利用词义注释的语料库，而基于知识的方法利用基础知识资源的结构和内容（例如定义或语义网络）。[8] 目前，监督方法明显优于基于知识的系统（Raganato、Camacho-Collados & Navigli，2017a）；但是，如前所述，它们严重依赖于通常稀缺的带有词义注释的语料库的可用性。

在本篇综述中，我们不会深入探讨WSD的更多细节。对于 WSD 的全面历史概述，我们推荐 Navigli (2009) 的综述，并且可以在 Raganato 等人（2017a）的经验比较中找到对当前方法的最新分析。

2.5 符号

在整篇综述中，我们使用以下符号。单词将被称为 w，而句子将被写为 s。概念、实体和关系将分别称为 c、e 和 r。在之前的工作 (Navigli, 2009) 之后，我们也使用以下可解释的词义表达：$word_n^p$ 是词性为 p 的词的第n个词义。至于在语义清单中表示的同义词集，我们将使用 y。[9] 语义网络通常表示为 N。为了引用向量，我们将在每个项目的顶部添加向量符号。例如，$\vec{w}$ 和$\vec{s}$将分别指代词 w 和意义 s 的向量。

一般来说，在本篇综述中，我们可以将语义表示称为一个通用的总称，包括超出单词级别的含义的所有向量表示（包括嵌入），或者明确特指与特定含义相关联的单词的向量表示[10]（例如，*bank* 它在经济学上的词义），无论它是否来自预定义语义清单，也不管它是指一个概念（例如，香蕉）还是一个实体（例如，法国）。

3. 无监督的词义表征

无监督语义表征仅基于从文本语料库中提取的信息构建。词义归纳，即自动识别词的可能含义，是这些技术的核心。无监督模型通过分析文本语料库中的上下文语义来归纳单词的不同含义，并根据从语料库中获得的统计知识来表示每种词义。根据模型使用的文本语料库的类型，我们可以将无监督的语义表示分为两大类：（1）仅利用单语语料库的技术（第 3.1 节）和（2）利用多语语料库的技术（第 3.2 节）。

3.1 利用单语语料库的词义表征

本节回顾使用未标记的单语语料库作为主要资源的感觉表示模型。这些方法可以分为两大类：（1）基于聚类（或两阶段）的模型（Van de Cruys, Poibeau, & Korhonen, 2011; Erk & Padó, 2008; Liu, Qiu, & Huang, 2015a），首先归纳词义，然后学习这些表征（第 3.1.1 节），以及（2）联合训练（Li & Jurafsky, 2015; Qiu, Tu, & Yu, 2016），它们一起执行归纳和表征学习（第 3.1 节。 2）。此外，在第 3.1.3 节中，我们将简要概述上下文嵌入，这是一个新兴的无监督技术分支，它从不同的角度看待语义表征。

3.1.1 两阶段模型

Schütze（1998）的语境-群体区分是语义表征的开创性工作之一。该方法是一种自动词义消歧的尝试，以解决语义注释数据的知识获取瓶颈（Gale 等，1992）和对外部资源的依赖。上下文组区分的基本思想是从上下文相似性中自动诱导词义，通过对出现歧义词的上下文进行聚类来计算。具体来说，歧义词w的每个上下文 C表示为一个词向量$\vec{v}_C$，计算为其内容词向量$\vec{v}_c$ $(c \in C)$的质心。为给定语料库中的每个单词计算上下文向量，然后使用期望最大化算法 (Dempster, Laird, & Rubin, 1977, EM) 将其聚类到预定数量的集群（上下文组）中。单词的上下文组被视为单词不同含义的表示。尽管很简单，但 Schütze (1998) 的基于聚类的方法构成了许多后续技术的基础，主要区别在于它们的上下文表示或底层聚类算法。图 4 描述了两阶段无监督感知表示技术所遵循的一般过程。

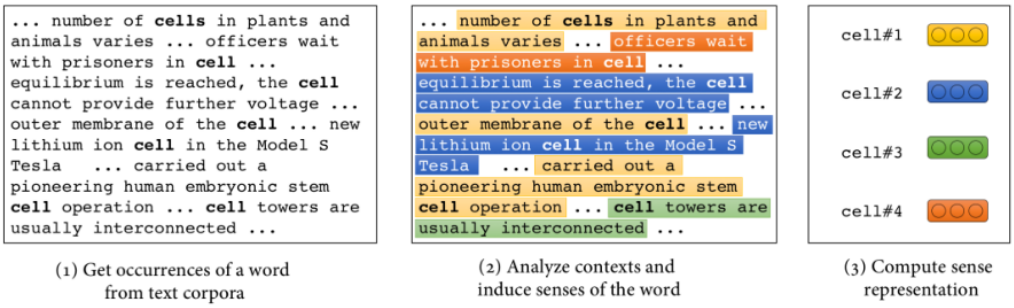


图 4：无监督语义表征技术首先归纳给定单词的不同词义（通常通过对该单词在文本语料库中的出现进行聚类），然后计算每个归纳词义的表征。

鉴于它需要为给定单词的所有单独上下文计算独立表示，上下文组区分方法不容易扩展到大型语料库。Reisinger 和 Mooney (2010) 通过直接对上下文进行聚类来解决这个问题，表示为 unigram 的特征向量，而不是将上下文建模为向量。该方法可以被认为是第一个新一代的语义表示技术，通常被称为多原型。在这项具体工作中，使用 von Mises-Fisher 分布混合 (movMF) 算法对上下文进行聚类。该算法类似于 k-means，但允许使用每个集群的浓度参数来控制语义广度，这将更好地模拟集群大小的倾斜分布。

类似地，Huang、Socher、Manning 和 Ng (2012) 提出了一种基于聚类的语义表征技术，具有三个不同之处： (1) 上下文向量是通过对它们的词向量进行 idf权重平均获得的； (2) 使用球形k-means进行聚类； (3) 最重要的是，一个词的首次出现用它们的簇标记，而第二遍用于学习语义表征。 Vu 和 Parker (2016) 也将两遍学习的想法用于另一种感觉表征建模架构。

语义表征也可以从语义网络中获得。例如，Pelevina、Arefyev、Biemann 和 Panchenko (2016) 通过将每个单词与其语义相似的单词集连接起来，构建了一个语义图。使用 Chinese Whispers 算法 (Biemann, 2006) 对这个网络中的节点进行聚类，并将语义作为每个聚类中单词的加权平均值来归纳。Sense-aware Semantic Analysis (Wu & Giles, 2015, SaSA) 采用了类似的语义归纳技术。 SaSA 遵循 Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007, ESA)，使用 Wikipedia 概念表示一个词。不是构建最近邻图，而是通过将所有相关文章收集到单词w并将它们聚类来构建 Wikipedia 文章图。然后在维基百科文章的语义空间上执行语义归纳步骤。

3.1.2 联合模型

基于聚类的语义表征方法受到聚类和语义表征彼此独立完成的限制，因此，这两个阶段没有利用它们固有的相似性。嵌入模型的引入是词义向量空间模型最具革命性的变化之一。作为一个密切相关的领域，语义表征并非不受影响。许多研究人员提出了 Skip-gram 模型的各种扩展 (Mikolov et al., 2013a)，这将能够捕获特定语义的区别。两阶段模型的一个主要限制是它们的计算成本高[11]。由于嵌入算法的效率及其统一的性质（与更传统技术的两阶段性质相反），这些技术通常是有效的。因此，许多最近的技术都依赖于嵌入模型作为它们的基础框架。

Neelakantan等人(2014) 是第一个提出 Skip-gram 模型的多原型扩展的人。他们的模型称为Multiple-Sense Skip-Gram (MSSG)，与早期的工作类似，它将单词的上下文表示为单词向量的质心，并将它们聚类以形成目标单词的语义表征。不过，根本区别在于聚类和语义嵌入学习是联合执行的。在训练期间，每个单词的预期词义被动态选择为最接近上下文的词义，并且仅针对该词义更新权重。在同时进行的工作中，Tian、Dai、Bian、Gao、Zhang、Chen 和 Liu (2014) 提出了一种基于 Skip-gram 的语义表征技术，该技术显著减少了 Huang 等人（2012）的模型的参数数量。在这种情况下，Skip-gram 模型中的词嵌入被替换为有限混合模型，其中每个混合对应于一个词的原型。该多原型 Skip-gram 模型的训练采用了 EM 算法。

Liu、Liu、Chua 和 Sun (2015b) 认为，上述技术的局限性在于它们仅考虑单词的局部上下文来归纳其语义表征。为了解决这个限制，他们提出了主题词嵌入（TWE），其中允许每个词在不同领域下具有不同的嵌入，其中领域是使用潜在领域建模全局计算的（Blei, Ng, & Jordan, 2003）。提出了模型的三个变体： (1) TWE-1，将每个领域视为一个伪词，分别学习主题嵌入和词嵌入； (2) TWE-2，将每个词领域视为一个伪词，直接学习TWE； (3) TWE-3，它为每个单词和每个领域分配不同的嵌入，并通过连接相

应的单词和领域嵌入来构建每个单词-领域对的嵌入。多种TWE模型的扩展已被提出。神经张量 Skip-gram (NTSG) 模型 (Liu et al., 2015a) 将领域建模的相同想法应用于语义表征，但引入了张量以更好地学习单词和领域之间的交互。另一个扩展是 MSWE (Nguyen, Nguyen, Modi, Thater, & Pinkal, 2017)，它认为在给定的上下文中可能会触发一个单词的多种意义，并通过混合权重替换 TWE 中最合适的词义的选择反映了该词在上下文中与多种词义的不同关联程度。

然而，这些联合无监督模型存在两个限制。首先，为了便于实施，大多数无监督意义表示技术假设每个单词有固定数量的词义。这种假设远非现实。单词往往具有高度不同的含义，从一个（单义的）到几十个。在给定的意义上，通常，大多数单词都是单义的。例如，WordNet 3.0 中大约 80% 的单词是单义的，只有不到 5% 的单词具有 3 种以上的意义。然而，歧义词往往在真实文本中更频繁地出现，这略微平滑了词在多义词中的高度倾斜分布。表 1 显示了 SemCor (Miller et al., 1993) 中单词类型的语义数量分布，SemCor 是最大的可用语义注释数据集之一，包含数千个单词的大约 235,000 个词义注释。偏态分布清楚地表明，自然文本中的词类型往往具有不同数量的意义，正如其他研究中所讨论的那样（Piantadosi，2014；Bennett，Baldwin，Lau，McCarthy，& Bond，2016；Pasini & Navigli，2018）。

# Senses	2	3	4	5	6	7	8	9	10	11	12	≥ 12
Nouns	22%	17%	14%	13%	9%	7%	4%	4%	3%	3%	1%	3%
Verbs	15%	16%	14%	13%	9%	7%	5%	4%	4%	3%	1%	9%
Adjectives	23%	19%	15%	12%	8%	5%	2%	3%	3%	1%	2%	6%

表 1：SemCor 数据集中每词词义数量的分布（修剪频率 < 10 的词）

其次，大多数无监督模型的一个共同点是，它们扩展了 Skipgram 模型，方法是将单词对其上下文的条件（如在原始模型中）替换为对预期意义的附加条件。然而，这些模型中的上下文词并没有消除歧义。因此，语义嵌入取决于其上下文的词嵌入。

在下文中，我们回顾了一些直接针对解决上述联合无监督模型的这两个限制的方法：

1. **动态多义** 语义表示模型的变化多义问题的直接解决方案是设置由外部语义清单定义的单词的词义数量。 Nieto Pina 和 Johansson (2015) 的 Skip-gram 扩展遵循此方法。然而，通过将外部词典作为基础材料，此方法受到两个主要限制。首先，该模型无法处理词典中未定义的单词。其次，该模型假设基础文本定义的词义区别与词典指定的词义区别相匹配，这可能不一定是正确的。换句话说，并非一个词的所有意义都可能出现在文本中，或者词典可能不会涵盖该词在基础文本中的所有不同预期词义。更好的解决方案将涉及从基础文本中动态归纳词义。这种模型首先在 Neelakantan 等人（2014）的非参数 MSSG (NP-MSSG) 系统中实现。该模型将 Meyerson (2001) 的在线非参数聚类程序应用于任务，仅当一个词类型与现有词义的相似度（使用当前上下文计算）小于参数λ时才为该词类型创建新词义。AdaGram (Bartunov, Kondrashkin, Osokin, & Vetrov, 2016) 通过更有原则的非参数贝叶斯方法改进了这种动态行为。该模型与之前的工作类似，建立在 Skip-gram 之上，假设一个词的多义词与其频率成正比（更频繁的词可能更具有多义词）。
2. **纯词义为基础的模型** Qie理想情况下，模型应该对词义选择之间的依赖关系进行建模，以解决上下文词的歧义。邱等人（2016）通过提出一个纯粹的基于词义的模型来解决这个问题。该模型还将消歧上下文从一个小窗口（如之前的工作中所做的那样）扩展到整个句子。 MUSE (Lee & Chen, 2017) 是另一个 Skip-gram 扩展，它使用强化学习提出纯词义表征。依赖于线性时间感知序列解码模块，该方法提供了一种搜索感知组合的更有效方式。

3.1.3 语境化词嵌入

鉴于无监督的语义表征通常是由聚类产生的，它们的语义区别不清楚，并且它们并不容易映射到明确定义的概念。事实上，这些模型的主要限制之一在于它们难以集成到下游模型中（有关这方面的更多详细信息，请参见第 6.2 节）。最近，一个新兴的研究分支专注于将无监督词嵌入直接集成到下游模型中。词嵌入，如 Word2vec 和 GloVe，为每个词计算一个单一的表示，用于表示下游模型中的单词，独立于它们出现的上下文。相比之下，上下文化的词嵌入对上下文很敏感，即它们的表征会根据它们出现的上下文动态变化。

Li 和 McCallum (2005) 的序列标注器是采用情景化表征的开创性工作之一。该模型基于软词聚类为每个词推断上下文敏感的潜在变量，并将它们作为附加特征集成到 CRF 序列标记器中。自 2011 年以来，随着词嵌入的引入（Collobert et al., 2011; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013c）和神经网络的功效，并且鉴于词嵌入的含义混淆缺陷，上下文相关模型再次引起研究关注。新兴解决方案主要旨在解决无监督技术的应用限制；因此，它们通常的特点是易于集成到下游应用程序中。 Context2vec (Melamud, Goldberger, & Dagan, 2016) 是上下文化表示新分支中最早和最突出的提议之一。该模型通过提取建立在双向 LSTM 语言模型之上的多层感知器的输出嵌入来表示目标词的上下文。 Context2vec 构成了许多后续工作的基础。

图 5 提供了将上下文化词嵌入集成到 NLP 模型中的高级说明。在训练时，对于给定输入文本中的每个单词（例如图中的*cell*），语言模型单元负责分析上下文（通常使用循环神经网络）并（调节）调整目标单词的表示它的上下文。这些上下文相关的嵌入实际上是深度循环神经网络的内部状态，无论是在单语语言建模设置中（Peters, Ammar, Bhagavatula, & Power, 2017; Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018) 或双语翻译配置 (McCann, Bradbury, Xiong, & Socher, 2017)。上下文嵌入的训练作为预训练阶段进行，独立于大型未标记或不同标记文本语料库的主要任务。在测试时，一个单词的上下文嵌入通常与其静态嵌入连接并馈送到主模型（Peters et al., 2018）。

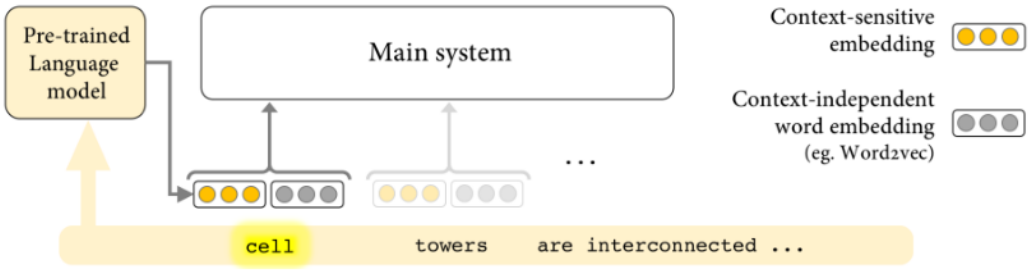


图 5：上下文化词嵌入的一般说明以及它们如何集成到 NLP 模型中（图中的*Main system*）。语言建模组件负责分析目标词（图中的*cell*）的上下文并生成其动态嵌入。与具有静态表示的（与上下文无关的）词嵌入不同，上下文化嵌入具有对其上下文敏感的动态表示。

Peters 等人 (2017) 的 TagLM 模型是该分支的一个最新示例，它在单语文本上训练多层双向 LSTM (Hochreiter & Schmidhuber, 1997) 语言模型。著名的 ELMo（语言模型嵌入）技术（Peters et al,. 2018 ）在原理上是相似的，只是在语言建模单元的两个方向之间共享了一些权重。 McCann 等人（2017）的上下文向量 (CoVe) 模型类似地使用两层双向 LSTM 网络计算上下文表示。但在机器翻译任务中，使用来自注意力序列到序列机器翻译模型的 LSTM 编码器对 CoVe 向量进行预训练。 [12]

3.2 利用多语言语料库的词义表征

由诸如 WordNet 之类的词义字典定义的词义差异对于某些下游应用程序（例如机器翻译 (MT)）可能不是最佳的。鉴于歧义不一定会跨语言转移，因此理想情况下，MT 的意义差异应根据特定语言对之间的翻译差异来定义。执行此操作的常用方法是对目标语言中源词的可能翻译进行聚类，每个聚类表示源词的特定含义。

这种翻译特定的意义清单已在 MT 文献中广泛使用（Ide, Erjavec, & Tufis, 2002; Carpuat & Wu, 2007b; Bansal, Denero, & Lin, 2012; Liu, Lu, & Neubig, 2018）。相同的差异可用于创建适合 MT 的语义嵌入。 Guo、Che、Wang 和 Liu (2014) 通过在平行语料库中对单词的翻译进行聚类，以同样的方式归纳了词义清单。源语言中的单词被标记了它们对应的词义，并且自动注释的数据被用来使用标准的词嵌入技术来计算词嵌入。 Ettinger、Resnik 和 Carpuat (2016) 遵循相同的词义归纳程序，但使用了 Jauhar、Dyer 和 Hovy (2015) [13] 的基于改进的词义表征，将原始模型 (WordNet) 中使用的标准词义库存替换为翻译特定库。

同样，Suster、Titov 和 van Noord (2016) 利用翻译差异作为自动编码器中的监督信号来归纳词义表征。在编码阶段，离散状态自动编码器为目标词分配意义，并在解码期间恢复给定词及其词义的上下文。在训练时，编码器使用单词及其翻译（来自对齐的语料库）。 Upadhyay、Chang、Zou、Taddy 和 Kalai (2017) 将这种双语模型扩展到多语言环境，以便更好地从多语言分布信息中受益。

4. 基于知识的语义表征

除了仅从文本语料库中学习的无监督技术之外，还有另一个研究分支利用了外部资源中可用的知识。本节介绍利用知识资源构建意义和概念表示的技术。首先，我们将概述当前使用的知识资源（第 4.1 节）。然后，我们将简要介绍一些利用知识资源改进词向量的方法（第 4.2 节）。最后，我们将专注于构建基于知识的词义表征（第 4.3 节）和概念或实体（第 4.4 节）。

4.1 知识资源

知识资源以多种形式存在。在本节中，我们概述了主要用于语义和概念表示学习的知识资源。知识资源的性质因几个因素而异。知识资源可以大致分为两大类：专家制作的和协作构建的。每种类型都有其自身的优点和局限性。专家制作的资源（例如 WordNet）具有准确的词典信息，例如文本定义、示例和概念之间的语义关系。另一方面，协作构建的资源（例如，维基百科或维基词典）提供了诸如百科全书式信息、更广泛的覆盖范围、多语言性和最新性等特性。[14]

在下文中，我们描述了词汇语义中用于表示学习的一些最重要的资源，即 WordNet（第 4.1.1 节）、维基百科和相关工作（第 4.1.2 节），以及不同资源的合并，例如 BabelNet 和概念网（第 4.1.3 节）。

4.1.1 WordNet

专家制造资源的一个突出例子是 **WordNet** (Miller, 1995)，它是 NLP 和语义表示学习中使用最广泛的资源之一。 WordNet 的基本组成部分是同义词集。同义词集代表一个独特的概念，可以通过名词、动词、形容词或副词来表达，并且由一个或多个词汇化（即用于表达该概念的同义词）组成。例如，定义为“形成骨骼轴并保护脊髓的一系列椎骨”概念的同义词包括六个词汇化：脊柱、脊柱、脊柱、脊椎、背部和轴。一个词可以属于多个同义词集，表示它可以采取的不同含义。因此，WordNet 也可以被视为词义字典。该词典中的词义定义在文献中广泛用于语义表征学习。

WordNet 也可以被视为一个语义网络，其中节点是同义词集，边是连接不同同义词集的词汇或语义关系（例如上位词或分词）。 WordNet 的最新版本（3.1，2012 年发布）涵盖 155,327 个单词和 117,979 个同义词集。在成为多语言资源的过程中，WordNet 还通过开放多语言 WordNet 项目（Bond & Foster，2013）和相关努力扩展到英语以外的语言。

4.1.2 维基百科、Freebase、维基数据和 DBpedia

协作构建的知识资源对包括 NLP 在内的广泛领域的研究做出了重大贡献。**维基百科**是此类资源最突出的例子之一。维基百科是世界上语言知识最大的多语种百科全书，有超过 250 种语言的数百万个概念和实体的单独页面。由于合作作者的不断更新，它的覆盖范围正在稳步增长。例如，仅英文维基百科每天就收到大约 750 篇新文章。每篇维基百科文章代表一个明确的概念（例如，*Spring*（设备））或实体（例如，*Washington*（州）），包含大量以文本信息、表格、信息框和各种关系（例如重定向）形式存在的信息，如消歧义和分类。

类似的合作成果是 **Freebase**（Bollacker、Evans、Paritosh、Sturge & Taylor，2008）。 Freebase 部分由 Wikipedia 提供支持，它是以知识库的形式收集大量结构化数据。截至 2014 年 1 月，Freebase 包含超过 4000 万个实体和 20 亿个关系。 Freebase 最终于 2016 年 5 月关闭，但其信息部分转移到 Wikidata 并用于构建 Google 的知识图谱。 **Wikidata** (Vrandecic, 2012) 是一个由维基媒体基金会直接运营的项目，旨在将维基百科转变为完全结构化的资源，从而提供可供其他维基媒体项目使用的通用数据源。它被设计为基于项目的面向文档的语义数据库，每个项目代表一个主题并由唯一标识符标识。知识以属性-值对形式的语句编码，其中还包括定义（描述）。 **DBpedia** (Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak, & Hellmann, 2009) 是构建维基百科内容的类似工作。特别是，DBpedia 利用了构成其主要信息来源的 Wikipedia 信息框。

4.1.3 BabelNet 和ConceptNet

基于专家和协作构建的资源中可用的知识类型使它们经常互补。这促使研究人员将这两个类别的各种词汇资源结合起来（Niemann & Gurevych, 2011; McCrae, Aguadode Cea, Buitelaar, Cimiano, Declerck, Gómez-Pérez, Gracia, Hollink, Montiel-Ponsoda, Spohr, et al., 2012； Pilehvar & Navigli, 2014）。一个突出的例子是 **BabelNet**（Navigli & Ponzetto，2012），它提供了 WordNet 与许多协作构建的资源（包括 Wikipedia）的合并。 BabelNet 的结构类似于 WordNet。同义词集是主要的语言单元，并与其他语义相关的同义词集相连，在这种情况下，它们的词汇化是多语言的。同义词集之间的关系来自 WordNet 以及来自其他资源（如 Wikipedia 超链接和 Wikidata）的新语义关系。这些资源的结合使 BabelNet 成为一个大型的多语言语义网络，在其 4.0 版本中包含 284 种语言的 15,780,364 个同义词集和 277,036,611 个词汇语义关系。

ConceptNet (Speer, Chin, & Havasi, 2017) 是一种类似的资源，它结合了来自异构来源的语义信息。特别是，ConceptNet 包括来自 WordNet、Wiktionary 和 DBpedia 等资源的关系，以及来自众包和有目的的游戏的常识性知识。 ConceptNet 和 BabelNet 之间的主要区别在于它们的主要语义单元：ConceptNet 对单词进行建模，而 BabelNet 使用 WordNet 样式的同义词集。

4.2 知识增强的词表示

如第 2 节所述，词向量表示（例如，词嵌入）主要是通过仅利用来自文本语料库的信息来构建的。然而，还有一系列研究试图将文本语料库中可用的信息与词汇资源中编码的知识相结合。可以利用这些知识来包含文本语料库中不可用的附加信息，以提高现有词向量表示的语义连贯性或覆盖率。此外，知识增强的词表示技术与基于知识的语义表示学习密切相关（见下一节），因为各种模型可以互换使用类似的技术。

早期使用词汇资源改进词嵌入的尝试修改了用于学习词嵌入的神经语言模型的目标函数（例如 Word2vec 的 Skip-gram），以便将来自词汇资源的关系整合到学习过程中（Xu, Bai, Bian ,Gao, Wang, Liu & Liu，2014； Yu & Dredze，2014 ）。最近的一类技术，通常称为改造（Faruqui、Dodge、Jauhar、Dyer、Hovy 和 Smith，2015 ），尝试通过后处理步骤改进预训练的词嵌入。给定任何预训练的词嵌入，改造的主要思想是移动通过给定语义网络中的关系连接更近的词。[15] 在改造模型中最小化的主要目标函数如下：

$$\sum_{i=1}^{|V|}\left(\alpha_i\left\|\vec{w}_i-\vec{\hat{w}}_i\right\|+\sum_{(w_i,w_j)\in \mathrm{N}}\beta_{i,j}\left\|\vec{w}_i-\vec{w}_j\right\|\right) \quad (2)$$

其中$|V|$表示词汇的大小，N是表示为一组词对的输入语义网络，$\vec{w}_i$和$\vec{w}_j$对应于预训练模型中的词嵌入，α_i和$\beta_{i,j}$是可调整的控制值，$\vec{\hat{w}}_i$表示输出词嵌入。

在改造的基础上，Speer 和 Lowry-Duda (2017) 利用 ConceptNet 的多语言关系信息在多语言空间上构建嵌入，Lengerich、Maas 和 Potts (2017) 通过显式建模成对关系来推广改造方法。其他类似的方法是 Pilehvar 和 Collier (2017) 以及 Goikoetxea、Soroa 和 Agirre (2015) 的方法，他们通过 Personalized Page Rank (Haveliwala, 2002) 分析语义网络的结构，分别以扩展预训练单词的覆盖范围和质量嵌入。最后，Bollegala、Alsuhaibani、Maehara 和 Kawarabayashi (2016) 修改了给定词嵌入模型的损失函数，通过同时利用来自共现和语义网络的线索来学习向量表示。

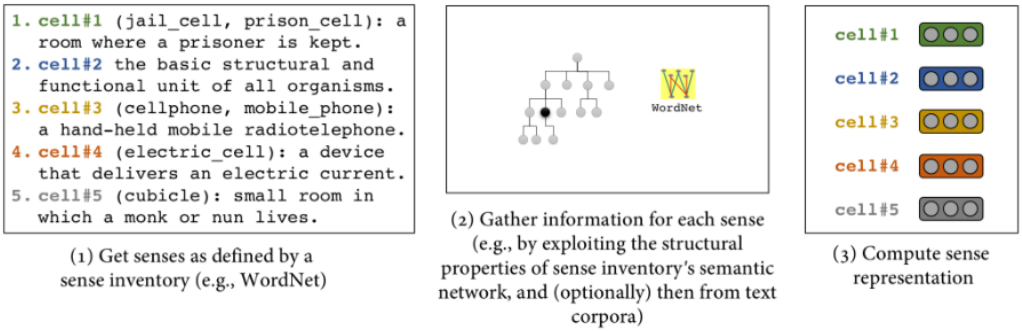
最近，出现了一个专注于专门针对特定应用的词嵌入的新分支。例如，Kiela、Hill 和 Clark (2015) 研究了两种改进的变体，以专门针对相似性或相关性的词嵌入，以及 Mrksic、Vulic、Séaghdha、Leviant、Reichart、Gai、Korhonen 和 Young (2017) 通过利用来自 PPDB 和 BabelNet 等资源的许多单语言和跨语言语言约束（例如，同义词和反义词），嵌入语义相似性和对话状态跟踪来专业化词向量。

事实上，正如最后一项工作所示，知识资源在多语言向量空间的构建中也发挥着重要作用。使用外部资源避免了编译大型并行语料库的需要，这在传统上一直是文献中学习跨语言词嵌入的主要来源（Upadhyay, Faruqui, Dyer, & Roth, 2016; Ruder, Vulic , & Søgaard, 2017）。这些用于学习跨语言嵌入的替代模型利用来自词汇资源的知识，例如 WordNet 或 BabelNet (Mrksic et al., 2017; Goikoetxea, Soroa, & Agirre, 2018)，双语词典 (Mikolov, Le, & Sutskever, 2013b; Ammar, Mulcaire , Tsvetkov,

Lample, Dyer, & Smith, 2016; Artetxe, Labaka, & Agirre, 2016; Doval, Camacho-Collados, Espinosa-Anke, & Schockaert, 2018) 或从维基百科中提取的类似语料库 (Vulic & Moens, 2015 ）。

4.3 基于知识的语义表征

本节概述了基于知识的语义表征的最新技术。这些表征通常是通过将一个词分解为其单独的词义表征而获得的，如外部词义清单所定义的那样。图 6 描述了基于知识的感知向量表征建模技术的主要工作流程。这些技术的学习信号各不相同，但主要利用了词汇资源中可用的两种不同类型的信息：文本定义（或注释）和语义网络。



(1) Get senses as defined by a sense inventory (e.g., WordNet)

(2) Gather information for each sense (e.g., by exploiting the structural properties of sense inventory's semantic network, and (optionally) then from text corpora)

(3) Compute sense representation

图 6：基于知识的语义表征技术对由外部词汇资源（意义清单）定义的单词进行词义区分。对于每种词义，都会收集相关信息并计算表征。

文本定义被用作通过几种方法初始化意义嵌入的主要信号。 Chen、Liu 和 Sun（2014 年）通过对在文本语料库上训练的预训练词嵌入进行平均，提出了词义嵌入的初始化。然后，这些初始化的词义表征用于消除大型语料库的歧义。最后，修改了 Word2vec (Mikolov et al., 2013a) 的 Skip-gram 的训练目标，以便从消歧的语料库中学习单词和语义嵌入。相比之下，Chen、Xu、He 和 Wang（2015 年）利用卷积神经网络架构使用来自词汇资源的文本定义来初始化语义嵌入。然后，这些初始化的语义嵌入被输入到 Neelakantan 等人 （2014）的多语义 Skip-gram 模型的一个变体中。（见第 3.1 节）用于学习基于知识的语义嵌入。最后，在 Yang 和 Mao (2016) 中，词义嵌入是通过在词对的短上下文中利用改编的 Lesk [16] 算法 (Vasilescu, Langlais, & Lapalme, 2004) 来学习的。

不同的研究方向已经尝试使用词汇资源的图形结构来学习基于知识的意义表示。正如 4.1 节所解释的，许多现有的词汇资源可以被视为**语义网络**，其中节点是概念，边表示概念之间的关系。语义网络构成了消除大量文本歧义的合适知识资源（Agirre et al., 2014; Moro et al., 2014）。因此，学习语义表征的一种直接方法是自动消除文本语料库的歧义，并对生成的带有语义注释的文本应用单词表示学习方法（Iacobacci, Pilehvar, & Navigli, 2015）。按照这个方向，Mancini、Camacho-Collados、Iacobacci 和 Navigli (2017) 提出了一种基于浅图的消歧程序，并修改了 Word2vec 的目标函数，以便在共享向量空间中同时学习词和词义嵌入。目标函数本质上类似于 Chen 等人 （2014）提出的目标函数。之前解释过，它还在学习过程的最后一步学习单词和语义嵌入。

与使用知识资源对词嵌入进行后处理类似（参见第 4.2 节），最近的工作利用预训练的词嵌入不仅可以改进它们，还可以将它们分解为词义。下面列出了**对预训练词嵌入**进行后处理以学习语义嵌入的方法：

1. 从语义网络中获得语义表征的一种方法是直接应用个性化 PageRank 算法 (Haveliwala, 2002)，如 Pilehvar 和 Navigli (2015) 所做的那样。该算法执行一组随机图游走，以计算每个 WordNet 同义词集（网络中的节点）的向量表示。使用类似的基于随机游走的程序，Pilehvar 和 Collier (2016) 为每个 WordNet 词提取了一组感觉偏差词。基于这些，他们提出了一种称为 DeConf 的方法，该方法将预先训练的词嵌入空间作为输入，并将一组语义嵌入（由 WordNet 定义）添加到同一空间。 DeConf 通过将一个词在空间中的嵌入推到其对应的词义偏词所占据的区域（对于词的特定意义）来实现这一点。图 7 显示了单词 *digit* 及其在向量空间中的诱发的*hand*和*number*意义。

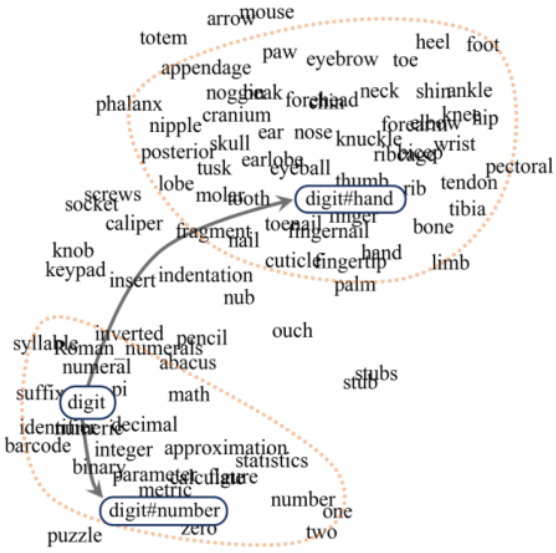


图 7：词和词义的混合语义空间。 DeConf (Pilehvar & Collier, 2016) 在词嵌入空间中引入了两个新点，用于词 digit 的数学和身体部分意义，从而产生混合空间。

2. Jauhar等人（2015 年） 提出了对改造[17]（Faruqui 等人，2015 年）的扩展，用于学习基础词义字典（例如 WordNet）的语义表征。他们还提出了第二种方法，该方法调整了 Word2vec 的训练目标，以在学习过程中包含语义。训练目标使用 EM 进行优化。

3. Johansson 和 Pina (2015) 通过具有两个主要约束的优化公式对预训练的词嵌入进行后处理：多义词嵌入可以分解为它们相应的意义嵌入的组合，并且意义嵌入应该靠近语义网络中的邻居。他们的实验中使用了瑞典语义网络 SALDO (Borin, Forsberg, & Lönngren, 2013)，尽管他们的方法也可以直接扩展到不同的语义网络。
4. 最后，AutoExtend (Rothe & Schütze, 2015) 是另一种使用预训练词嵌入作为输入的方法。在这种情况下，他们提出了一种基于两个主要约束的自动编码器架构：一个词向量对应于其意义向量的总和，一个同义词集对应于其词汇化（意义）的总和。例如，单词*crane*的向量将对应于其词义 $crane_n^1$、$crane_n^2$ 和$crane_v^1$的向量之和（使用WordNet 作为参考）。类似地，在 WordNet 中定义为“提前安排和保留（给别人的东西）”的同义词集的向量将等于其对应的语义 *Reserve*、*hold* 和 *book* 的向量之和。等式 3 以数学方式显示了这些约束：

$$\vec{w} = \sum_{i=1}^{n} \vec{s}_i; \quad \vec{y} = \sum_{j=1}^{m} \vec{s}_j \quad (3)$$

其中$\vec{s}_i$和$\vec{s}_j$分别指代词w和同义词集y的意义。

4.4 概念和实体表征

在本节中，我们介绍了仅依赖于知识库的关系信息（第 4.4.1 节）和结合了来自文本语料库和知识资源的线索的混合模型（第 4.4.2 节）的方法。

4.4.1 知识库嵌入

本节对仅从知识库中针对概念和命名实体的那些表征技术进行回顾。该领域的大量研究将知识图（或语义网络）作为信号来构建实体（和关系）的表示，专门针对知识库完成任务[18]。

该领域的一项开创性工作是 TransE (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013)，一种嵌入实体和关系的方法。在这个模型中，关系被视为在与实体相同的向量空间中运行的翻译。给定一个表示为一组三元组 $(e1, r, e2)$的知识库，其中 $e1$ 和 $e2$ 是实体，r 是它们之间的关系，主要目标是对于空间中的所有三元组（即，$\forall(e1, r, e2) \in N$ ）以 $\vec{e}_1 + \vec{r} \approx \vec{e}_2$ 的方式接近实体。图 8 说明了模型背后的主要思想。这个目标可以通过利用不同的学习架构和约束来实现。在 Bordes 等人（2013）的

原创工作中，优化是通过随机梯度下降进行的，其中嵌入的 L2 归一化作为附加约束。遵循这一基本思想，各种方法提出了对学习架构不同部分的改进：

1. TransP (Wang, Zhang, Feng, & Chen, 2014b) 是一个类似的模型，它通过处理知识图中存在的特定属性来改进关系映射。
2. Lin、Liu、Sun、Liu 和 Zhu (2015) 提出在不同空间中学习实体和关系的嵌入 (TransR)。
3. Ji, He, Xu, Liu, and Zhao (2015) 介绍了分离空间中每个实体关系对的动态映射 (TransD)。
4. Luo、Wang、Wang 和 Guo (2015) 提出了一种使用预训练词嵌入进行初始化的两阶段架构。
5. Yang、Yih、He、Gao 和 Deng（2015 年）提出了一个概括 TransE 和 NTN 的统一学习框架（Socher、Perelygin、Wu、Chuang、Manning、Ng 和 Potts，2013 年）。
6. 最后，Ebisu 和 Ichise (2018) 讨论了 TransE 的正则化问题并提出了 TorusE，它受益于解决 TransE 正则化问题的新正则化方法。

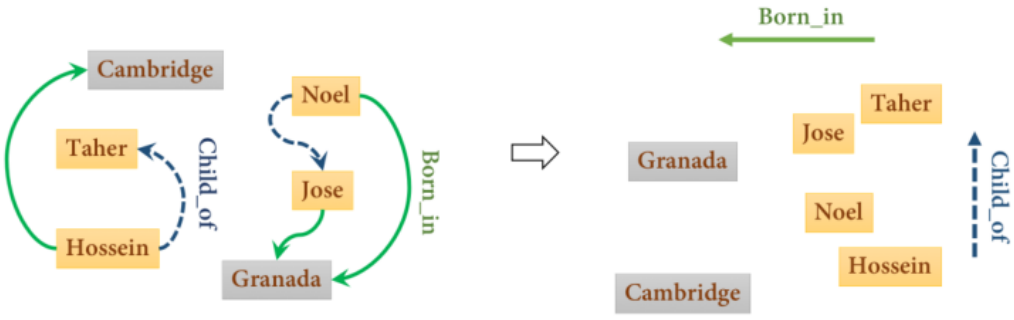


图 8：从知识图到实体和关系嵌入。插图创意基于 Weston 和 Bordes (2014) 的幻灯片。

或者，一个研究分支专门专注于建模实体（而不是关系）并计算图中各个节点的嵌入。 DeepWalk (Perozzi, Al-Rfou, & Skiena, 2014) 是该分支中的突出技术之一。该算法的核心思想是使用随机图游走将给定图表示为一系列人工句子。与语义相似的词倾向于共现的自然语言类似，这些人工句子中的连续词对应于图中的相邻（拓扑相关）顶点。然后将这些句子用作 Skip-gram 模型的输入（参见第 2.1.2 节），并计算单个单词（即概念节点）的嵌入。 Node2vec (Grover & Leskovec, 2016) 是 DeepWalk 的扩展，它更好地控制了随机游走的深度优先和广度优先属性。相比之下，Nickel 和 Kiela (2017) 通过将单词嵌入到庞加莱球中提出了一种新的表示形式[19]，它考虑了作为输入给出的分类的相似性和层次结构[20]。

这些是近年来关于知识库嵌入的一些最相关的工作，但鉴于有关该主题的大量论文，这篇评论绝不是全面的。 Cai、Zheng 和 Chang (2018) 或 Nguyen (2017) 对知识图嵌入进行了更广泛的概述，包括更深入的解释，后者侧重于知识库完成任务。

4.4.2 利用知识库和文本语料库的混合模型

除了完全依赖于知识库中可用信息的技术之外，还有一些模型将来自知识库和文本语料库的线索组合到相同的表示中。鉴于其半结构化性质和所提供的文本内容，维基百科一直是此类表示的主要来源。虽然大多数方法都使用维基百科注释的语料库作为学习维基百科概念和实体表示的主要来源（Wang, Zhang, Feng, & Chen, 2014a; Sherkat & Milios, 2017; Cao, Huang, Ji, Chen, & Li , 2017），(Camacho-Collados, Pilehvar, & Navigli, 2016)[21] 还探索了来自 Wikipedia 和 WordNet 等异构资源的知识组合 。

鉴于它们的混合性质，这些模型也可以很容易地用于文本应用程序。一个简单的应用是词或命名实体消歧，其中嵌入可用作神经网络架构嵌入层的初始化（Fang, Zhang, Wang, Chen, & Li, 2016; Eshel, Cohen, Radinsky, Markovitch, Yamada, & Levy, 2017) 或直接用作利用语义相似性的基于知识的消歧系统 (Camacho-Collados et al., 2016)。

5. 评价

在本节中，我们介绍了用于评估意义表示质量的最常见的评估基准。根据其性质，评估程序通常分为内在（第 5.1 节）和外在（第 5.2 节）。

5.1 内在评价

内在评估是指一类基准，它提供对向量空间的质量和连贯性的通用评估，独立于它们在下游应用程序中的性能。可以从本质上测试不同的属性，语义相似性传统上被视为评估意义表示的最直接的特征。尤其是不需要组合性的词和词组等小词汇单元的语义相似性受到了最多的关注。单词相似度数据集以多种形式存在。区分相似性和相关性的概念也存在重要的价值。虽然语义上相似的词可以在上下文中相互替换，但相关词足以在相同的上下文中（例如，在文档中）同时出现，而不需要可替换性。 WordSim-353 (Finkelstein, Evgeniy, Yossi, Ehud, Zach, Gadi, & Eytan, 2002) 是一个将这两个概念混为一谈的数据集。真正的相似性数据集包括仅包含 65 个词对的 RG-65 (Rubenstein & Goodenough, 1965)，或由 999 个词对组成的 SimLex-999 (Hill, Reichart, & Korhonen, 2015)。此外，还有多语言基准，其中包括多种语言的单词相似度数据集。例如，WordSim-353 和 SimLex-999 (Leviant & Reichart, 2015) 的翻译和再注释以及来自 SemEval-2017 多语言单词相似度任务的数据集 (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017)英语以外语言的评估基准。

为了使这些基于词的评估基准适应语义向量，已经提出了各种策略（Reisinger & Mooney，2010）。其中，最流行的是在两个词之间取最相似的一对词义（Resnik，1995；Pilehvar & Navigli，2015；Mancini et al.，2017），也称为 MaxSim：

$$\operatorname{sim}\left(w_1, w_2\right)=\max_{s_1 \in S_{w_1}, s_2 \in S_{w_2}} \cos\left(\vec{s}_1, \vec{s}_2\right)$$

其中S_{wi} 是一个包含所有w_i词义的集合，$\vec{s}_i$表示词语s_i的语义向量表征。另一种策略，称为 *AvgSim*，简单地平均 w_1 和 w_2 所有可能意义的成对相似性。余弦相似度（cos）是计算意义向量之间相似度的最突出的度量。

在所有这些基准测试中，单词都是单独配对的。但是，我们知道要触发一个模棱两可的词的特定含义，该词需要出现在特定的上下文中。事实上，Kilgarriff (1997) 认为，用一组固定的词义表示一个词可能不是建模词义的最佳方法，而是应该根据给定的上下文定义词义。为此，Huang（2012）等人提出了一种不同类型的相似性数据集，其中为单词提供了相应的上下文。该任务包括通过考虑两个词出现的上下文来评估两个词的相似性。该数据集被称为斯坦福语境词相似度（SCWS），并已被确立为意义表示的主要内在评估之一。需要一个预消歧步骤来利用此任务中的语义表征。通常使用简单的相似性度量，例如 *MaxSimC* 或 *AvgSimC*。与 *MaxSim* 和 *AvgSim* 不同，*MaxSimC* 和 *AvgSimC* 将目标词的上下文考虑在内。首先，计算在句子中选择最合适意义的置信度（例如，通过计算上下文中词嵌入的平均值，并根据余弦相似度选择最接近平均上下文向量的语义）。然后，最终得分对应于所选词义之间的相似性（即 *MaxSimC* ）或所有词义之间的加权平均值（即 *AvgSimC* ）。

然而，即使在这个数据集上，语义表征通常优于基于单词的模型，用于消除输入文本歧义的简单策略可能并不是最优的。事实上，最近的研究表明，使用 *AvgSim* 在单词相似性任务中基于语义的模型的改进可能不是由于准确的含义建模，而是由于相关的人工工作，例如子采样，这些人工工作(Dubossarsky, Grossman, & Weinshall，2018）没有得到控制 。这与最近的一项研究一致，该研究分析了感知和语境化表征在语境中捕捉意义的能力（Pilehvar & Camacho-Collados，2018）。该分析中提出的二进制分类任务包括确定目标词在两个不同上下文中的出现是否对应于相同的含义。结果表明，最近的语义[22]和上下文化表示技术无法准确区分上下文中的含义，其性能仅比简单的基线略好，同时显着落后于数据集的人类评估者间一致性[23]。

最后，除了这些任务之外，还有其他内在评估程序，例如同义词选择 (Landauer & Dumais, 1997; Turney, 2001; Jarmasz & Szpakowicz, 2003; Reisinger & Mooney, 2010)、异常值检测 (Camacho-Collados & Navigli , 2016; Blair, Merhav, & Barry, 2016; Stanovsky & Hopkins, 2018) 或感觉聚类 (Snow, Prakash, Jurafsky, & Ng, 2007; Dandala, Hokamp, Mihalcea, & Bunescu, 2013)。有关更多信息，Bakarov (2018) 提供了对内在评估基准的更全面概述。

5.2 外在评价

外部评估程序旨在评估下游任务中含义表征的质量。除了内在评估程序外，外在评估对于理解不同意义表示技术在实际应用中的有效性也是必要的。这一点尤其重要，因为内在评估协议并不总是与下游性能相关（Tsvetkov, Faruqui, Ling, Lample, & Dyer, 2015; Chiu, Korhonen, & Pyysalo, 2016; Faruqui, Tsvetkov, Rastogi, & Dyer, 2016）。然而，虽然外部评估对于评估在下游任务中集成感觉表示的有效性绝对重要，但与更直接的内部程序相比，任务、流水线和基准的可变性也更高。

在自然语言处理中用作语义表征的外部评估程序的一些最常见的任务是文本分类和情感分析（Liu et al., 2015b; Li & Jurafsky, 2015; Pilehvar, Camacho-Collados, Navigli, & Collier, 2017)、文档相似性 (Wu & Giles, 2015) 和词义归纳 (Pelevina et al., 2016; Panchenko, Ruppert, Faralli, Ponzetto, & Biemann, 2017b) 和消歧 (Chen et al., 2014; Rothe & Schütze，2015；Camacho-Collados et al.，2016；Peters et al.，2018）。如第 4.4.1 节所述，知识库嵌入也经常在知识库完成任务上进行评估（Bordes et al., 2013）。在接下来的部分中，我们将更详细地解释一些迄今为止已经应用了意义表示的应用程序。

6. 应用

如第 5 节和整篇综述中所述，语义表征研究的主要目标之一是使这些知识载体能够有效地集成到下游应用程序中。与单词表示（更具体地说是嵌入）不同，语义表征在这方面仍处于起步阶段。这也是由于这些表示的非立即集成，这通常需要额外的词义消歧或归纳步骤。然而，与词嵌入一样，语义表征理论上可以应用于多种应用。

将语义表征集成到下游应用程序中并不是一个新趋势。自九十年代以来，针对重要的基于文本的应用程序在这个方向上出现了许多不同的努力，并取得了不同程度的成功。**信息检索**是研究词义集成的首批应用之一。在较早的一项尝试中，Schütze 和 Pedersen (1995) 展示了基于词义的文档查询相似性如何导致基于词的模型的显着改进。

机器翻译 (MT) 是另一个见证了不断努力整合语义级信息的经典任务。由于一个词可能会根据其在上下文中的预期含义而具有不同的翻译，因此传统上认为意义识别能够潜在地改进基于词的 MT 模型。卡Carpuat等人分析了 WSD 对当时标准 MT 系统性能的影响（Carpuat & Wu, 2005, 2007a, 2007b）。这些研究尚无定论，但普遍反映了将基于语义的模型成功集成到 MT 流水线中的困难。这也部分是由于缺乏意义注释的语料库，产生了知识获取瓶颈（Gale et al., 1992）[24]。

从那时起，词义（尤其是语义表征）已被集成到各种 NLP 任务中。在下文中，我们将讨论意义表示（第 6.1 节）和最近的上下文化表示（第 6.2 节）在下游任务中的应用。

6.1 语义表征的应用

文献中将无监督的语义表征集成到下游应用程序中是有限的。 Li 和 Jurafsky (2015) 提出了一个框架，将无监督的语义嵌入集成到各种自然语言处理任务中。该研究得出的结论是，所提出的无监督表征没有提供显着的影响，这表明词嵌入维数的增加可以导致类似的结果。然而，消歧步骤是一个基于语义嵌入和输入文本的嵌入表示之间的相似性（计算为内容词嵌入的平均值）的简单过程。最近的一个提议是 Kartsaklis、Pilehvar 和 Collier (2018) 的 Multi-Sense LSTM 模型，它避免了明确消歧的需要。该系统在神经网络中用 k个单独的语义嵌入替换了每个单词的传统单词嵌入层。对于每个训练实例，使用注意力机制动态选择预期语义并相应更新。根据单词出现的上下文，在测试时使用类似的机制。该系统在各种任务中被证明是有效的，在文本到实体映射的多个基准测试中报告了最先进的性能。

就基于知识为基础的表征而言，需要一个显式或隐式的词义消歧步骤来将词转换为它们的预期语义。Pilehvar 等人（2017）提出了一种基于单词和基于知识的语义嵌入共享空间的方法，在将它们集成到用于文本分类的神经网络架构之前引入了一个简单的基于图的消歧步骤。当输入文本足够大时，包含的语义会有所改善，但在此设置中包含预训练的语义嵌入并没有显着改善大多数数据集中词嵌入的使用。在使用超感时观察到使用语义表征的主要好处（参见第 7.3 节），Flekova 和 Gurevych (2016) 在其他下游分类任务中也观察到了这一结论。

除了这些研究之外，还有其他应用可以有效地整合更广泛意义上的语义和概念表示：词义或命名实体消歧（Chen et al., 2014; Rothe & Schütze, 2015; CamachoCollados et al., 2016; Fang et al., 2016; Panchenko, Faralli, Ponzetto, & Biemann, 2017a; Peters et al., 2018），知识库完成 (Bordes et al., 2013) 或统一 (Delli Bovi, Espinosa- Anke, & Navigli, 2015), 常识推理 (Lieto, Radicioni, Rho, & Mensa, 2017), 词汇替换 (Cocos, Apidianaki, & Callison-Burch, 2016), 上位词发现 (Espinosa-Anke, Camacho-Collados , Delli Bovi, & Saggion, 2016），词汇蕴涵 (Nickel & Kiela, 2017），或视觉对象发现 (Young, Kunze, Basile, Cabrio, Hawes, & Caputo, 2017)。

6.2 情境化表征的应用

上下文化表征（参见第 3.1.3 节）通过采用不同的策略，为必须将输入句子离散化为词义的问题提供了另一种解决方案。在这种情况下，输入词没有明确地替换为语义嵌入；但是，它们的表示是根据上下文动态调整的（因此，隐式消歧）。

由于它们的动态特性，上下文化词嵌入可以无缝集成到神经架构中，因此，它们已经在广泛的 NLP 任务中进行了评估，包括情感分析、问答和分类、文本蕴涵、语义角色标签、阅读理解、命名实体提取和共指解析（Peters et al., 2018; Salant & Berant, 2018; McCann et al., 2017）。上下文对应物替换传统的静态词嵌入的改进，证明了具有可以根据其上下文调整目标词的语义的动态表示的优势。其他最近的例子包括 Wanxiang Che 和 Liu (2018) 的 HIT-SCIR 系统，该系统在 CoNLL 2018 通用依赖解析共享任务中获得了最佳性能（Zeman、Hajic、Popel、Potthast、Straka、Ginter、Nivre 和 Petrov , 2018) 通过采用 ELMo 嵌入 (Peters et al., 2018) 和 Liu 等人（2018）的端到端神经机器翻译架构，它使用上下文感知嵌入显式地对同形异义词（即歧义词）进行建模，从而提高了歧义词的翻译性能。

7. 分析

本节对基于知识的表示技术和无监督表示技术进行了分析和比较，强调了每种技术的优点和局限性，同时提出了每种技术适合的设置和场景。我们关注四个重要方面：可解释性（第 7.1 节）、对不同领域的适应性（第 7.2 节）、感觉粒度（第 7.3 节）和组合性（第 7.4 节）。

7.1 可解释性

从词到词义层次的主要原因之一是词义的语义基础性质，这可能会带来更好的可解释性。然而，在这个特定方面，无监督模型和基于知识的模型之间存在相当大的差异。无监督模型直接从文本语料库中学习语义，从而产生特定于模型的意义解释。这些诱发的词义不一定与人类的词义区别概念相对应，或者不容易区分。出于这个原因，已经提出了一些方法来提高无监督感觉表示的可解释性，或者通过提取它们的上位词或它们的视觉表示（即，说明特定含义的图像）（Panchenko 等人，2017b）或通过映射诱导的感觉到外部感官清单（Panchenko，2016）。

相比之下，基于知识的表征已经与语义清单中的条目相关联，这使得可解释性更高，因为这些条目通常与定义、示例、图像相关联，并且通常与其他概念（例如 WordNet）和翻译（例如，BabelNet）。反过来，这可以从词汇资源中直接注入额外的先验信息，这可能有助于为终端模型提供更深入的背景知识（Young et al., 2017）。作为一个缺点，基于知识的表征通常受限于潜在的意义清单，因此可能无法准确表示文本语料库中不存在的新语义。这可以通过保持更新感知清单来部分解决，尽管通常不是一个简单的过程。如第 4.1 节所述，像维基百科这样的协作资源不太容易受到这个问题的影响。

7.2 对不同领域的适应性

在词嵌入中受到称赞的一个特征是它们对一般和专业领域的适应性（Goldberg，2016）。从这方面来看，无监督模型比基于知识的模型具有理论优势，因为它们能够直接从给定的文本语料库中诱导语义。这使他们有机会根据手头的领域和给定的任务来调整他们的语义区分。相反，基于知识的系统通常会学习语义清单给出的所有语义的表示。因此，他们无法将自己的语义区分专门化到领域或使他们的粒度适应任务。

Mancini 等人（2017）提出的知识增强方或Fang等人（2016）直接从文本语料库中学习，可以部分缓解基于知识的模型的这种限制。然而，语义应该仍然存在于用作模型输入的语义网络中。换句话说，基于知识的方法无法学习新的语义，这可能是某些特定领域和任务的重要限制。此外，某些领域的准确表示将需要合适的知识资源，这可能不适用于专业领域或低资源语言。

7.3 感觉粒度

语义清单可能会列出几十种不同的词义，例如*run*、*play*和*get*。多义词（即歧义词）通常分为两类：多义词和同音词。多义词有多个相关的含义。例如，单词*mark*可以指"区别符号"以及"表面上的可见指示"。在这种情况下，这两种意义的区别也被称为细粒度的，因为这两种意义很难分开。同音词[25]具有完全不相关的含义。例如，*bank*[26]一词的地质和金融机构含义。这也是一种粗粒度区分的情况，因为*bank*的这两个含义明显不同。

一般来说，一些语义清单的细粒度一直是 NLP 中的一个争论点（Kilgarriff，1997；Navigli，2009；Hovy、Navigli 和 Ponzetto，2013）。有人指出，WordNet 中的语义区分可能过于细化，无法用于许多 NLP 应用程序（Navigli，2006；Snow et al.，2007；Hovy et al.，2013）。例如，WordNet 3.0（参见第 4.1.1 节）列出了动词 *run* 的 41 种不同含义。然而，这些意义中的大多数都被翻译成西班牙语的*correr* 或*operar*。因此，诸如机器翻译之类的多语言任务可能不会从语义清单提供的额外区别中受益。事实上，将这些细粒度的区别合并到更粗粒度的类（在 WordNet 中称为超感知）已被证明在各种下游应用中是有益的（Flekova & Gurevych，2016；Pilehvar et al.，2017）。

这个讨论也与无监督技术有关。语义的动态学习，而不是固定所有单词的词义数量，已经证明可以提供更真实的语义分布（参见第 3.1.2 节）。此外，已经讨论过是否所有出现的单词都可以有效地划分为词义（Kilgarriff，1997；Hanks，2000；Kilgarriff，2007），从而产生了一种以分级方式描述单词含义的新方案（ Erk et al.，2009；MeCarthy et al.，2016）。虽然本篇综述未涵盖该计划，但已表明评估语义的分级量表可能与人类感知不同语义的方式更好地相关。尽管得出的结论并不完全相同，但这些发现也与对当前感测库存细粒度的批评有关，这已被证明在某些下游应用中是有害的。

7.4 语意合成性

组合方法根据其组成部分（例如单词）的含义对复杂表达式的语义进行建模。通常，组成词被表示为它们的词向量，所有的含义都混合在一起。然而，对于表达式中的歧义词，通常只触发单一含义，其他含义无关紧要。因此，将单词的含义精确到给定的上下文可能是组合性的合理想法。这对于查询歧义可能成为问题的信息检索等应用至关重要（Allan & Raghavan, 2002; Di Marco & Navigli, 2013）。

不同的工作试图在组合性的背景下引入语义表征（Köper & im Walde, 2017; Kober, Weeds, Wilkie, Reffin, & Weir, 2017），取得了不同程度的成功。主要思想是选择一个单词的预期意义，并仅通过基于上下文的意义归纳 (Thater, Fürstenau, & Pinkal, 2011) 或基于示例的表示 (Reddy, Klapaftis, McCarthy) 将特定意义引入到工作中, & Manandhar, 2011)，或借助外部资源，例如 WordNet (Gamallo & Pereira-Farina, 2017)。 Cheng 和 Kartsaklis (2015) 中可以找到第一种方法的示例，其中提出了一种递归神经网络，其中词嵌入被分成多个意义向量。该网络应用于释义检测并取得了积极的结果。

一般来说，在组合性背景下对语义区分模型的评估通常在通用基准上进行评估，例如释义检测。尽管在诸如问答和信息检索等任务中具有潜在的好处，但没有尝试将语义表征整合为神经组合模型的组件。

8. 结论

在本篇综述中，我们对用于构建分布式意义表示的基于语义的模型进行了广泛的概述。词嵌入已被证明可以提供有趣的语义属性，可以应用于大多数语言应用程序。但是，这些模型倾向于将不同的含义混为一个表示。因此，要深入理解词汇意义，通常需要准确区分词义。为此，在本文中，我们讨论了学习语义表征的模型，这些语义表示可以直接从文本语料库（即无监督）或由外部语义清单定义（即基于知识）定义。

其中一些模型已经在实践中证明是有效的，但仍有很大的改进空间。例如，尽管几乎所有模型都（不同程度地）捕获了基于语义的信息，但常识推理尚未得到深入探索。此外，这些模型中的大多数仅在英语上进行了测试，而只有少数提出了其他语言的模型或尝试了多语言。最后，将这些理论模型整合到下游应用程序中是下一步，因为目前尚不清楚最佳整合策略是什么，以及是否需要预消歧步骤。例如，Peters 等人 （2018）的上下文嵌入等方法已经展示了一个新的可能方向，在该方向上，可以为每个上下文动态学习语义，而无需明确的预消歧步骤。

虽然不是完全分布式的语义表征，但以灵活的方式建模关系也是未来工作的另一种可能途径。关系通常在针对基于知识的完成的作品中建模。此外，最近的一项研究集中在借助文本语料库改进关系嵌入（Toutanova、Chen、Pantel、Poon、Choudhury & Gamon，2015；Jameel、Bouraoui & Schockaert，2018；Espinosa-Anke & Schockaert，2018），这为将这些关系集成到下游文本应用程序的新方法铺平了道路。

从这个角度来看，语义的定义和正确的范式当然仍然是一个悬而未决的问题。语义需要离散吗？它们是否需要与知识资源或感知库存相关联？是否应该根据上下文动态学习它们？这些是许多关于该主题的研究尚待探索的问题。正如我们在分析中所解释的那样，一些方法更适合某些应用程序或领域，没有任何明确的一般结论。这些悬而未决的问题当然仍然具有相关性，并鼓励对语义的分布式表示进行进一步研究，还有许多领域有待探索。

致谢

作者要感谢匿名审稿人的意见，他们的意见有助于提高本次调查的整体质量。 Jose Camacho-Collados 的研究得到 ERC 拨款 637277 的支持。

参考文献

1. Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, *40* (1), 57–84.
2. Allan, J., & Raghavan, H. (2002). Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 307–314, Tampere, Finland.
3. Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016).
4. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.
5. Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294.
6. Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pp. 789–798.
7. Athiwaratkun, B., & Wilson, A. (2017). Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1645–1656.
8. Athiwaratkun, B., Wilson, A., & Anandkumar, A. (2018). Probabilistic FastText for multi- sense word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–11. Association for Computational Linguistics.
9. Bakarov, A. (2018). A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.
10. Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley Framenet Project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 86–90. Association for Computational Linguistics.
11. Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for Word Sense Disambiguation using WordNet. In *Proceedings of the Third International Conference on*

Computational Linguistics and Intelligent Text Processing, CICLing'02, pp. 136–145, Mexico City, Mexico.

12. Bansal, M., Denero, J., & Lin, D. (2012). Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 773–782, Stroudsburg, PA, USA. Association for Computational Linguistics.
13. Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Vol. 51 of *Proceedings of Machine Learning Research*, pp. 130–138, Cadiz, Spain. PMLR.
14. Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Lan- guage Model. *The Journal of Machine Learning Research*, *3*, 1137–1155.
15. Bennett, A., Baldwin, T., Lau, J. H., McCarthy, D., & Bond, F. (2016). Lexsemtm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Pro- ceedings of ACL*, pp. 1513–1524.
16. Biemann, C. (2006). Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pp. 73–80. Association for Computational Linguistics.
17. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, *7* (3), 154–165.
18. Blair, P., Merhav, Y., & Barry, J. (2016). Automated generation of multilingual clusters for the evaluation of distributed representations. *arXiv preprint arXiv:1611.01547*.
19. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.
20. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, *5* (1), 135–146.
21. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collab- oratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. ACM.
22. Bollegala, D., Alsuhaibani, M., Maehara, T., & Kawarabayashi, K.-i. (2016). Joint word representation learning using a corpus and a semantic lexicon. In *AAAI*, pp. 2690– 2696.
23. Bond, F., & Foster, R. (2013). Linking and extending an open multilingual Wordnet. In
24. *ACL (1)*, pp. 1352–1362.
25. Bordes, A., Chopra, S., & Weston, J. (2014). Question answering with subgraph embeddings.
26. In *EMNLP*.
27. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pp. 2787–2795.
28. Borin, L., Forsberg, M., & L¨onngren, L. (2013). Saldo: a touch of yin to WordNets yang.
29. *Language resources and evaluation*, *47* (4), 1191–1211.
30. Cai, H., Zheng, V. W., & Chang, K. (2018). A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*.
31. Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A frame- work for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 43–50.
32. Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26.
33. Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities.

Artificial Intelligence, *240*, 36–64.
34. Cao, Y., Huang, L., Ji, H., Chen, X., & Li, J. (2017). Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1623–1633.
35. Carpuat, M., & Wu, D. (2005). Word sense disambiguation vs. statistical machine trans- lation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 387–394. Association for Computational Linguistics.
36. Carpuat, M., & Wu, D. (2007a). How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. *Proceedings of TMI*, 43–52.
37. Carpuat, M., & Wu, D. (2007b). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Meth- ods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
38. Chaplot, D. S., & Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. In *Proceedings of AAAI*.
39. Chen, T., Xu, R., He, Y., & Wang, X. (2015). Improving distributed representation of word sense via WordNet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing – Short Papers*, pp. 15–20, Beijing, China.
40. Chen, X., Liu, Z., & Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pp. 1025–1035, Doha, Qatar.
41. Cheng, J., & Kartsaklis, D. (2015). Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1531–1542. Association for Computa- tional Linguistics.
42. Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, Berlin, Germany.
43. Cocos, A., Apidianaki, M., & Callison-Burch (2016). Word sense filtering improves embedding-based lexical substitution. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 99–104. Association for Computational Linguistics.
44. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pp. 160–167.
45. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, *12*, 2493– 2537.
46. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & J´egou, H. (2018). Word translation without parallel data. In *Proceedings of ICLR*.
47. Dandala, B., Hokamp, C., Mihalcea, R., & Bunescu, R. C. (2013). Sense clustering using Wikipedia.. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 164–171, Hissar, Bulgaria.
48. Delli Bovi, C., Camacho-Collados, J., Raganato, A., & Navigli, R. (2017). EuroSense: Auto- matic harvesting of multilingual sense annotations from parallel text. In *Proceedings of ACL*, Vol. 2, pp. 594–600.
49. Delli Bovi, C., Espinosa-Anke, L., & Navigli, R. (2015). Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of EMNLP*, pp. 726–736. Association for Computational Linguistics.
50. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incom- plete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.
51. Di Marco, A., & Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, *39* (3), 709–754.
52. Doval, Y., Camacho-Collados, J., Espinosa-Anke, L., & Schockaert, S. (2018). Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 294–304.

53. Dubossarsky, H., Grossman, E., & Weinshall, D. (2018). Coming to your senses: on controls and evaluation sets in polysemy research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1732–1740, Brussels, Belgium.
54. Ebisu, T., & Ichise, R. (2018). Toruse: Knowledge graph embedding on a lie group. In
55. *Proceedings of the AAAI Conference on Artificial Intelligence*.
56. Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*, Prague, Czech Republic.
57. Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey.
58. *Language and Linguistics Compass*, *6* (10), 635–653.
59. Erk, K., McCarthy, D., & Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 10–18. Association for Computational Linguistics.
60. Erk, K., & Pad´o, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 897–906.
61. Eshel, Y., Cohen, N., Radinsky, K., Markovitch, S., Yamada, I., & Levy, O. (2017). Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Com- putational Natural Language Learning (CoNLL 2017)*, pp. 58–68.
62. Espinosa-Anke, L., Camacho-Collados, J., Delli Bovi, C., & Saggion, H. (2016). Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP*, pp. 424–435.
63. Espinosa-Anke, L., & Schockaert, S. (2018). Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2653–2665.
64. Ettinger, A., Resnik, P., & Carpuat, M. (2016). Retrofitting sense-specific word vectors using parallel text. In *Proceedings of NAACL-HLT*, pp. 1378–1383, San Diego, California.
65. Fang, W., Zhang, J., Wang, D., Chen, Z., & Li, M. (2016). Entity disambiguation by knowledge and text jointly embedding. In *Proceedings of CoNLL*, pp. 260–269.
66. Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, pp. 1606–1615.
67. Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 30–35. Association for Com- putational Linguistics.
68. Finkelstein, L., Evgeniy, G., Yossi, M., Ehud, R., Zach, S., Gadi, W., & Eytan, R. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20* (1), 116–131.
69. Flekova, L., & Gurevych, I. (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of ACL*.
70. Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pp. 1606–1611, Hyderabad, India.
71. Gale, W. A., Church, K., & Yarowsky, D. (1992). A method for disambiguating word senses in a corpus. *Computers and the Humanities*, *26*, 415–439.
72. Gamallo, P., & Pereira-Farin˜a, M. (2017). Compositional semantics using feature-based models from wordnet. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 1–11, Valencia, Spain. Association for Computational Linguistics.
73. Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pp. 758–764.
74. G¨ardenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.
75. Goikoetxea, J., Soroa, A., & Agirre, E. (2015). Random walks and neural network language models on knowledge bases. In *Proceedings of NAACL*, pp. 1434–1439.

76. Goikoetxea, J., Soroa, A., & Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*.
77. Goldberg, Y. (2016). A primer on neural network models for natural language processing.
78. *Journal of Artificial Intelligence Research*, *57*, 345–420.
79. Grover, A., & Leskovec, J. (2016). Node2Vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Dis- covery and Data Mining*, KDD '16, pp. 855–864, New York, NY, USA.
80. Guo, J., Che, W., Wang, H., & Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*, pp. 497–507.
81. Hanks, P. (2000). Do word meanings exist?. *Computers and the Humanities*, *34* (1-2), 205–215.
82. Harris, Z. (1954). Distributional structure. *Word*, *10*, 146–162.
83. Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, pp. 517–526, Hawaii, USA.
84. Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.
85. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Compasutation*, *9* (8), 1735–1780.
86. Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42* (1), 177–196.
87. Hovy, E. H., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, *194*, 2–27.
88. Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word represen- tations via global context and multiple word prototypes. In *Proceedings of ACL*, pp. 873–882, Jeju Island, Korea.
89. Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pp. 95–105, Beijing, China.
90. Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of ACL*, pp. 897–907, Berlin, Germany.
91. Ide, N., Erjavec, T., & Tufis, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of ACL-02 Workshop on WSD: Recent Successes and Future Directions*, pp. 54–60, Philadelphia, USA.
92. Jameel, S., Bouraoui, Z., & Schockaert, S. (2018). Unsupervised learning of distributional relation vectors. In *Proceedings of ACL*, Melbourne, Australia.
93. Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 212–219, Borovets, Bulgaria.
94. Jauhar, S. K., Dyer, C., & Hovy, E. (2015). Ontologically grounded multi-sense repre- sentation learning for semantic vector space models. In *Proceedings of NAACL*, pp. 683–693, Denver, Colorado.
95. Ji, G., He, S., Xu, L., Liu, K., & Zhao, J. (2015). Knowledge graph embedding via dy- namic mapping matrix. 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)*, Vol. 1, pp. 687–696.
96. Johansson, R., & Pina, L. N. (2015). Embedding a semantic network in a word space. In
97. *Proceedings of NAACL*, pp. 1428–1433, Denver, Colorado.
98. Jones, M. P., & Martin, J. H. (1997). Contextual spelling correction using latent seman- tic analysis. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pp. 166–173.
99. Kartsaklis, D., Pilehvar, M. T., & Collier, N. (2018). Mapping text to knowledge graph entities using multi-sense LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.
100. Khodak, M., Risteski, A., Fellbaum, C., & Arora, S. (2017). Automated WordNet construc- tion using word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity*

Representations and their Applications, pp. 12–23.
101. Kiela, D., Hill, F., & Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2044–2048.
102. Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities*, *31* (2), 91–113.
103. Kilgarriff, A. (2007). Word senses. In *Word Sense Disambiguation*, pp. 29–46. Springer.
104. Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pp. 1746–1751, Doha, Qatar.
105. Kober, T., Weeds, J., Wilkie, J., Reffin, J., & Weir, D. (2017). One representation per word - does it make sense for composition?. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 79–90, Valen- cia, Spain. Association for Computational Linguistics.
106. K¨oper, M., & im Walde, S. S. (2017). Applying multi-sense embeddings for german verbs to determine semantic relatedness and to detect non-literal language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2, pp. 535–542.
107. Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *SIGMOD Rec.*, *31* (2), 84–93.
108. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104* (2), 211.
109. Landauer, T., & Dooley, S. (2002). Latent semantic analysis: theory, method and applica- tion. In *Proceedings of CSCL*, pp. 742–743.
110. Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE software*, *14* (2), 67–75.
111. Lee, G.-H., & Chen, Y.-N. (2017). Muse: Modularizing unsupervised sense embeddings. In
112. *Proceedings of EMNLP*, Copenhagen, Denmark.
113. Lengerich, B. J., Maas, A. L., & Potts, C. (2017). Retrofitting distributional embeddings to knowledge graphs with functional relations. *arXiv preprint arXiv:1708.00112*.
114. Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual Conference on Systems Documentation,* Toronto, Ontario, Canada, pp. 24–26.
115. Leviant, I., & Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.
116. Levy, O., & Goldberg, Y. (2014a). Dependency-based word embeddings. In *ACL*, pp.
117. 302–308.
118. Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization.
119. In *Advances in neural information processing systems*, pp. 2177–2185.
120. Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding?. In *Proceedings of EMNLP*, pp. 683–693, Lisbon, Portugal.
121. Li, W., & McCallum, A. (2005). Semi-supervised sequence modeling with syntactic topic models. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, pp. 813–818. AAAI Press.
122. Lieto, A., Radicioni, D., Rho, V., & Mensa, E. (2017). Towards a unifying framework for conceptual represention and reasoning in cognitive systems. *Intelligenza Artificiale*, *11* (2), 139 –153.
123. Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion.. In *Proceedings of AAAI*, pp. 2181–2187.
124. Liu, F., Lu, H., & Neubig, G. (2018). Handling homographs in neural machine translation.
125. In *Proceedings of NAACL*, New Orleans, LA, USA.

126. Liu, P., Qiu, X., & Huang, X. (2015a). Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1284–1290.
127. Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015b). Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2418–2424.
128. Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28* (2), 203– 208.
129. Luo, F., Liu, T., Xia, Q., Chang, B., & Sui, Z. (2018). Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2473–2482.
130. Luo, Y., Wang, Q., Wang, B., & Guo, L. (2015). Context-dependent knowledge graph embedding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1656–1661.
131. Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9* (Nov), 2579–2605.
132. Mallery, J. C. (1988). *Thinking about foreign policy: Finding an appropriate role for ar- tificial intelligence computers, Ph.D. Thesis*. M.I.T. Political Science Department, Cambridge, MA.
133. Mancini, M., Camacho-Collados, J., Iacobacci, I., & Navigli, R. (2017). Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of CoNLL*, pp. 100–111, Vancouver, Canada.
134. McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems 30*, pp. 6294–6305. Curran Associates, Inc.
135. McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability.
136. *Computational Linguistics*.
137. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., G´omez-P´erez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, *46* (4), 701–719.
138. Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61, Berlin, Germany.
139. Meyerson, A. (2001). Online facility location. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pp. 426–432, Washington, DC, USA. IEEE Computer Society.
140. Mihalcea, R., & Csomai, A. (2007). Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge man- agement*, pp. 233–242, Lisbon, Portugal.
141. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*.
142. Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
143. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.
144. Mikolov, T., Yih, W.-t., & Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751.
145. Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, *38* (11), 39–41.
146. Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pp. 303– 308, Plainsboro, N.J.
147. Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*

(TACL), *2*, 231–244.
148. Mrksic, N., Vuli´c, I., S´eaghdha, D. O´., Leviant, I., Reichart, R., Gai, M., Korhonen, A., & Young, S. (2017). Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics (TACL)*.
149. Navigli, R. (2006). Meaningful clustering of senses helps boost Word Sense Disambiguation performance. In *Proceedings of COLING-ACL*, pp. 105–112, Sydney, Australia.
150. Navigli, R. (2009). Word Sense Disambiguation: A survey. *ACM Computing Surveys*, *41* (2), 1–69.
151. Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.
152. Neale, S. (2018). A Survey on Automatically-Constructed WordNets and their Evaluation: Lexical and Word Embedding-based Approaches. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., & Tokunaga, T. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
153. Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pp. 1059–1069, Doha, Qatar.
154. Nguyen, D. Q., Nguyen, D. Q., Modi, A., Thater, S., & Pinkal, M. (2017). A mixture model for learning multi-sense word embeddings. In *Proceedings of *SEM 2017*.
155. Nguyen, D. Q. (2017). An overview of embedding models of entities and relationships for knowledge base completion. *arXiv preprint arXiv:1703.08098*.
156. Nickel, M., & Kiela, D. (2017). Poincar´e embeddings for learning hierarchical representations. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 6341–6350. Curran Associates, Inc.
157. Niemann, E., & Gurevych, I. (2011). The people's web meets linguistic knowledge: au- tomatic sense alignment of Wikipedia and Wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 205–214.
158. Nieto Pin˜a, L., & Johansson, R. (2015). A simple and efficient method to generate word sense representations. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 465–472, Hissar, Bulgaria.
159. Otegi, A., Aranberri, N., Branco, A., Hajic, J., Neale, S., Osenova, P., Pereira, R., Popel, M., Silva, J., Simov, K., & Agirre, E. (2016). QTLeap WSD/NED Corpora: Semantic Annotation of Parallel Corpora in Six Languages. In *Proc. of LREC*, pp. 3023–3030.
160. Panchenko, A. (2016). Best of both worlds: Making word sense embeddings interpretable.
161. In *Proceedings of LREC*, pp. 2649–2655.
162. Panchenko, A., Faralli, S., Ponzetto, S. P., & Biemann, C. (2017a). Using linked disam- biguated distributional networks for word sense disambiguation. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 72–78.
163. Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., & Biemann, C. (2017b). Un- supervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of EACL*, pp. 86–98.
164. Pasini, T., & Navigli, R. (2018). Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of AAAI*, New Orleans, United States.
165. Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2016). Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 174–183.
166. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pp. 1532–1543.

167. Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 701–710.
168. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL*, New Orleans, LA, USA.
169. Peters, M., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1756– 1765. Association for Computational Linguistics.
170. Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21* (5), 1112–1130.
171. Pilehvar, M. T., & Camacho-Collados, J. (2018). Wic: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*.
172. Pilehvar, M. T., Camacho-Collados, J., Navigli, R., & Collier, N. (2017). Towards a Seamless Integration of Word Senses into Downstream NLP Applications. In *Proceedings of ACL*, Vancouver, Canada.
173. Pilehvar, M. T., & Collier, N. (2016). De-conflated semantic representations. In *Proceedings of EMNLP*, pp. 1680–1690, Austin, TX.
174. Pilehvar, M. T., & Collier, N. (2017). Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 388–393, Valencia, Spain. Association for Computational Linguistics.
175. Pilehvar, M. T., & Navigli, R. (2014). A robust approach to aligning heterogeneous lexical resources. In *Proceedings of ACL*, pp. 468–478.
176. Pilehvar, M. T., & Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, *228*, 95–128.
177. Pratt, L. Y. (1993). Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems 5*, pp. 204–211.
178. Qiu, L., Tu, K., & Yu, Y. (2016). Context-dependent sense embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 183–191.
179. Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 337–346.
180. Raganato, A., Camacho-Collados, J., & Navigli, R. (2017a). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL*, pp. 99–110, Valencia, Spain.
181. Raganato, A., Delli Bovi, C., & Navigli, R. (2017b). Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1156–1167.
182. Reddy, S., Klapaftis, I. P., McCarthy, D., & Manandhar, S. (2011). Dynamic and static prototype vectors for semantic composition. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pp. 705–713.
183. Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word mean- ing. In *Proceedings of ACL*, pp. 109–117.
184. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy.
185. In *Proceedings of IJCAI*, pp. 448–453.
186. Rothe, S., & Schu¨tze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pp. 1793–1803, Beijing, China.
187. Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8* (10), 627–633.
188. Ruder, S. (2017). On word embeddings, part 1. *URL: http://ruder.io/word-embeddings-2017/(visited on 1/04/2018)*.

189. Ruder, S., Vuli´c, I., & Søgaard, A. (2017). A survey of cross-lingual word embedding models.
190. *arXiv preprint arXiv:1706.04902*.
191. Salant, S., & Berant, J. (2018). Contextualized word representations for reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 554–559, New Orleans, Louisiana.
192. Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw- Hill, New York.
193. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing.
194. *Communications of the ACM*, *18* (11), 613–620.
195. Schu¨tze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, *24* (1), 97–123.
196. Schu¨tze, H., & Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of SDAIR'95*, pp. 161–175, Las Vegas, Nevada.
197. Schwartz, R., Reichart, R., & Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 258–267.
198. Sherkat, E., & Milios, E. E. (2017). Vector embedding of Wikipedia concepts and enti- ties. In *International Conference on Applications of Natural Language to Information Systems*, pp. 418–428. Springer.
199. Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to merge word senses.
200. In *Proceedings of EMNLP*, pp. 1005–1014, Prague, Czech Republic.
201. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Parsing with compositional vector grammars. In *Proceedings of EMNLP*, pp. 455–465, Sofia, Bulgaria.
202. Soucy, P., & Mineau, G. W. (2005). Beyond TFIDF weighting for text categorization in the vector space model. In *Proceedings of IJCAI*, Vol. 5, pp. 1130–1135.
203. Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4444–4451.
204. Speer, R., & Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th Inter- national Workshop on Semantic Evaluation (SemEval-2017)*, pp. 85–89. Association for Computational Linguistics.
205. Stanovsky, G., & Hopkins, M. (2018). Spot the odd man out: Exploring the associative power of lexical resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
206. S˘uster, S., Titov, I., & van Noord, G. (2016). Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*, pp. 1346–1356, San Diego, California.
207. Taghipour, K., & Ng, H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. *CoNLL 2015*, 338.
208. Thater, S., Fu¨rstenau, H., & Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1134–1143, Chiang Mai, Thailand.
209. Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., & Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pp. 151–160.
210. Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499– 1509.
211. Tripodi, R., & Pelillo, M. (2017). A game-theoretic approach to word sense disambiguation.
212. *Computational Linguistics*, *43* (1), 31–70.
213. Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP (2)*, pp. 2049–2054, Lisbon,

Portugal.
214. Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 384–394, Uppsala, Sweden.
215. Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In
216. *Proceedings of the 12th European Conference on Machine Learning*, pp. 491–502.
217. Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pp. 417–424.
218. Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.
219. Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psycho- logical Review*, *89* (2), 123.
220. Upadhyay, S., Chang, K.-W., Zou, J., Taddy, M., & Kalai, A. (2017). Beyond bilingual: Multi-sense word embeddings using multilingual context. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada.
221. Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1661–1670.
222. Ustalov, D., Panchenko, A., & Biemann, C. (2017). Watset: Automatic induction of synsets from a graph of synonyms. In *Proceedings of the 55th Annual Meeting of the Associ- ation for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1579–1590.
223. Van de Cruys, T., Poibeau, T., & Korhonen, A. (2011). Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1012–1022, Edinburgh, Scotland, UK.
224. Vasilescu, F., Langlais, P., & Lapalme, G. (2004). Evaluating variants of the lesk approach for disambiguating words. In *Proceddings of LREC*.
225. Vilnis, L., & McCallum, A. (2015). Word representations via Gaussian embedding. In
226. *Proceedings of ICLR*.
227. Vrandeˇci´c, D. (2012). Wikidata: A New Platform for Collaborative Data Collection. In
228. *Proceedings of WWW*, pp. 1063–1064.
229. Vu, T., & Parker, D. S. (2016). K-embeddings: Learning conceptual embeddings for words using context. In *Proceedings of NAACL-HLT*, pp. 1262–1267.
230. Vuli´c, I., & Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*
231. *Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2, pp. 719–725.
232. Vyas, Y., & Carpuat, M. (2017). Detecting asymmetric semantic relations in context: A case-study on hypernymy detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pp. 33–43.
233. Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014a). Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1591–1601.
234. Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014b). Knowledge graph embedding by trans- lating on hyperplanes. In *Proceedings of AAAI*, pp. 1112–1119.
235. Wanxiang Che, Yijia Liu, Y. W. B. Z., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.
236. Weaver, W. (1955). Translation. *Machine translation of languages*, *14*, 15–23.
237. Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured training for neural network transition-based parsing. In *Proceedings of ACL*, pp. 323–333, Beijing, China.
238. Weston, J., & Bordes, A. (2014). Embedding methods for NLP. In *EMNLP Tutorial*.

239. Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Em- pirical Methods in Natural Language Processing*, pp. 1504–1515.
240. Wu, Z., & Giles, C. L. (2015). Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia. In *AAAI*, pp. 2188–2194. Citeseer.
241. Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., & Liu, T.-Y. (2014). Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1219–1228. ACM.
242. Yaghoobzadeh, Y., & Schu¨tze, H. (2016). Intrinsic subspace evaluation of word embedding representations. In *Proceedings of ACL*, pp. 236–246.
243. Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of ICLR*.
244. Yang, X., & Mao, K. (2016). Learning multi-prototype word embedding from single- prototype word embedding with integrated knowledge. *Expert Systems with Applica- tions*, *56*, 291 – 299.
245. Young, J., Kunze, L., Basile, V., Cabrio, E., Hawes, N., & Caputo, B. (2017). Semantic web-mining and deep vision for lifelong object discovery. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 2774–2779. IEEE.
246. Yu, M., & Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In
247. *ACL (2)*, pp. 545–550.
248. Yuan, D., Richardson, J., Doherty, R., Evans, C., & Altendorf, E. (2016). Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING*, pp. 1374– 1385.
249. Zeman, D., Haji˘c, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov,
250. S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–21, Brussels, Belgium. Association for Computational Linguistics.
251. Zhong, Z., & Ng, H. T. (2010). It Makes Sense: A wide-coverage Word Sense Disambiguation system for free text. In *Proceedings of the ACL System Demonstrations*, pp. 78–83, Uppsala, Sweden.
252. Zipf, G. K. (1949). *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.
253. Zou, W. Y., Socher, R., Cer, D. M., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pp. 1393–1398, Seattle, USA.

---

1. 指甲也可以指布料测量单位（通常是一码的十六分之一），甚至可以用作动词。

2. 为了获得目标词的可能词义列表，词典编纂者倾向于首先从语料库中收集词的出现，然后根据上下文手动对它们进行语义聚类，即一致性（Kilgarriff，1997）。鉴于此过程，Kilgarriff (1997) 建议，由 NLP 中的意义清单定义的词义不应被解释为对象，而应被解释为对词用法集群的抽象。

3. 例如，WordNet 3.0 中的 155K 单词中约有 83% 被列为单义词（有关词汇资源的更多信息，请参见第 4.1 节）。

4. 有关词嵌入及其当前挑战的更全面概述，请参阅 Ruder (2017) 的综述。

5. 使用 PCA 降低维度；使用 http://projector.tensorflow.org/ 进行可视化。

6. 在第 4.1 节中，我们提供了几个最流行的语义清单的概述。

7. 在这篇综述中，我们还涵盖了与知识资源直接相关的表征，即使没有明确列出词义（例如，维基百科中的概念和实体），包括知识库嵌入。

8. 一些方法也可以归类为混合方法，因为它们同时利用了语义注释的语料库和知识资源，例如，Luo 等人的gloss-augmented模型。（2018 年）。

9. 有关这些资源相关概念的更多信息，请参见第 4.1 节。

10. 在某些作品中，语义也被称为词位（Rothe & Schütze，2015）。

11. 例如，Huang 等人的模型。(2012) 花了大约一周的时间在 10 亿个标记的语料库上学习 100,000 个词汇中的 6,000 个子集的语义嵌入 (Neelakantan et al., 2014)。

12. 一般来说，上下文嵌入的预训练属性使其与迁移学习密切相关（Pratt，1993），这超出了本文的范围。

13. 有关此模型的更多详细信息，请参见第 4.3 节。

14. 除了这两种类型的资源之外，最近的另一个分支正在研究从头开始自动构建知识资源（尤其是 WordNet 类）（Khodak, Risteski, Fellbaum & Arora, 2017；Ustalov, Panchenko & Biemann, 2017）。但是，这些输出资源尚未在实践中使用，并且已被证明通常缺乏召回率（Neale，2018）。

15. FrameNet (Baker, Fillmore, & Lowe, 1998)、WordNet 和 PPDB (Ganitkevitch, Van Durme, & CallisonBurch, 2013) 用于他们的实验。

16. 最初的 Lesk 算法 (Lesk, 1986) 及其变体利用文本定义和目标词上下文之间的相似性来消除歧义。

17. 有关改装的更多信息，请参见第 4.2 节。

18. 给定一个不完整的知识库作为输入，知识库完成包括预测原始资源中缺失的关系的任务。

19. 庞加莱球是一个双曲空间，其中所有点都在单位圆内。

20. WordNet 在原始工作中用作参考分类法。

21. Wikipedia 和 WordNet 的结合依赖于 BabelNet 提供的多语言映射（有关 BabelNet 的更多信息，请参见第 4.1.3 节）。

22. 与 *MaxSimC* 技术类似，通过检索最接近基于上下文的向量的意义嵌入来评估语义表征，通过平均其词嵌入来计算。

23. 另一项将上位词检测任务作为他们实验的试验台的研究（Vyas & Carpuat，2017）得出了类似的结论。

24. 值得一提的是，虽然仍然缺乏有语义注释的多语言语料库，但最近的努力已经通过（半）自动消除大量并行语料库的歧义直接解决了这个问题（Taghipour & Ng, 2015; Otegi, Aranberri, Branco , Hajic, Neale, Osenova, Pereira, Popel, Silva, Simov, & Agirre, 2016；Delli Bovi, Camacho-Collados, Raganato, & Navigli, 2017）。

25. 根据剑桥词典，同音异义词是"与另一个词发音相同（同音字）或拼写相同（同形异义词）但含义不同的词"。鉴于 NLP 侧重于书面形式，在这种情况下，同音异义词通常是指后一种情况，即具有不同含义的同形异义词。

26. 同音异义词和多义词之间的区别有时可能很微妙。例如，历史语言学的研究表明，bank这个词的两个含义在意大利语中可能是相互关联的，因为银行家过去常常在河岸上做生意。