A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs

嵌入式空间对齐方法综述 语言和知识图表

Alexander Kalinowski
亚历山大 • 卡林诺夫斯基
College of Computing and
Informatics Drexel University
德雷克塞尔大学计算与信息学院
Philadelphia, PA 19104
宾夕法尼亚州费城，邮编 19104
ajk437@drexel.edu
ajk437@drexel.edu

Yuan An
袁安
College of Computing
and Informatics Drexel
University
德雷克塞尔大学计算
与信息学院
Philadelphia, PA 19104
宾夕法尼亚州费城，
邮编 19104
ya45@drexel.edu
ya45@drexel.edu

October 27, 2020
2020 年 10 月 27 日

ABSTRACT
摘要

Neural embedding approaches have become a staple in the fields of computer vision, natural language processing, and more recently, graph analytics. Given the pervasive nature of these algorithms, the natural question becomes how to exploit the embedding spaces to map, or align, embeddings of different data sources. To this end, we survey the current research landscape on word, sentence and knowledge graph embedding algorithms. We provide a classification of the relevant alignment techniques and discuss benchmark datasets used in this field of research. By gathering these diverse approaches into a singular survey, we hope to further motivate research into alignment of embedding spaces of varied data types and sources.

神经嵌入方法已经成为计算机视觉、自然语言处理以及最近的图形分析领域的主要内容。考虑到这些算法的普遍性，自然的问题就变成了如何利用嵌入空间来映射或对齐不同数据源的嵌入。为此，我们综述了单词、句子和知识图嵌入算法的研究现状。我们提供了相关比对技术的分类，并讨论了该研究领域中使用的基准数据集。通过将这些不同的方法收集到一个单一的调查中，我们希望进一步激励对不同数据类型和来源的嵌入空间的对齐的研究。

Keywords Knowledge Graph · Language Models · Entity Alignment · Cross-Lingual Alignment
知识图；语言模型；实体对齐；跨语言对齐

1 Introduction

2 介绍

The purpose of this survey is to explore the core techniques and categorizations of methods for aligning low-dimensional embedding spaces. Projecting sparse, high-dimensional data sets into compact, lower-dimensional spaces allows not only for a significant reduction in storage space, but also builds dense representations with many applications. These embedding spaces have become a staple in representation learning ever since their heralded application to natural language in a technique called word2vec, and have replaced traditional machine learning features as easy-to-build, high-quality representations of the source objects. There has been a wealth of study around techniques for embedding objects, such as images, natural language and knowledge graphs, and many research agendas focused on mapping one embedding space to another, either for the purpose of aligning and unifying to a common space, applications to joint downstream tasks or ease of transfer learning. In order to fully leverage these dense representations and translate them across domains and problem spaces, techniques for establishing alignments between them must be developed and understood. To this extent, we believe this avenue of research will continue to blossom, motivating us to present this survey of current methods for alignment of representations of both text and knowledge graphs, classifying them into a taxonomy based on their mathematical approaches and requisite levels of parallel data sources.

这项调查的目的是探索低维嵌入空间对齐方法的核心技术和分类。将稀疏的高维数据集投影到紧凑的低维空间中，不仅可以显著减少存储空间，还可以构建许多应用程序的密集表示。自从被称为 word2vec 的技术应用于自然语言以来，这些嵌入空间已经成为表示学习的主要内容，并取代了传统的机器学习功能，成为源对象的易于构建的高质量表示。围绕嵌入对象(例如图像、自然语言和知识图)的技术已经有了大量的研究，并且许多研究议程

集中于将一个嵌入空间映射到另一个嵌入空间，或者是为了对齐和统一到公共空间、应用于联合下游任务或者是为了迁移学习的方便。为了充分利用这些密集的表示，并跨领域和问题空间翻译它们，必须开发和理解在它们之间建立比对的技术。在这种程度上，我们相信这种研究途径将继续蓬勃发展，这促使我们提出对文本和知识图表示对齐的当前方法的调查，根据它们的数学方法和并行数据源的必要级别将它们分类。

Before diving into a survey of alignment techniques, let us first consider their motivation. In today's age, the availability of data is pervasive. In machine learning, however, methods for learning from this data are skewed to those of supervised learning where labeled data instances are required to optimize models and generalize to new applications. Labeled data is a bottleneck for the majority of machine learning projects in industry, and while methods for eliminating this bottleneck have been proposed with degrees of success [1], the luxury of labeled data is not always available without prohibitive cost.

在深入调查比对技术之前，让我们首先考虑它们的动机。在当今时代，数据的可用性无处不在。然而，在机器学习中，从这些数据中学习的方法倾向于监督学习的方法，在监督学习中，需要标记的数据实例来优化模型并推广到新的应用。标记数据是工业中大多数机器学习项目的瓶颈，尽管已经提出了消除这一瓶颈的方法并取得了一定程度的成功[1]，没有高昂的成本，标记数据的奢侈并不总是可用的。

At the same time, methods for learning a lower-dimensional representation of data in an unsupervised way have proven useful as inputs to machine learning algorithms. These representation learning algorithms, or embeddings, have become a de-facto approach for generating dense and compact feature sets, eliminating the need for tedious human engineering

同时，以无监督的方式学习数据的低维表示的方法已经被证明是作为机器学习算法的输入是有用的。这些表示学习算法或嵌入已经成为产生密集和紧凑特征集的事实上的方法，消除了对乏味的人类工程的需要

of features at the onset of every new task. The success of these techniques is not only related to their proven accuracy in downstream tasks, but their ability to train without supervision, thereby allowing them to scale to massive datasets.

在每个新任务开始的时候。这些技术的成功不仅与它们在下游任务中被证明的准确性有关，还与它们在没有监督的情况下进行训练的能力有关，从而允许它们扩展到大规模数据集。

Given the lack of available supervised data and the prevalence of strong unsupervised methods for embedding large datasets, we consider the problem of matching between embeddings of one set to another, henceforth referred to as embedding alignment. We believe that alignment methods provide a convenient method for deriving correspondences between pairs of embedding spaces, thereby providing a method for bootstrapping fuzzy labels between the source and target spaces. This problem has applications in, but not limited to, the following areas.

考虑到缺乏可用的监督数据和用于嵌入大数据集的强非监督方法的流行，我们考虑一个集合到另一个集合的嵌入之间的匹配问题，以下称为嵌入对齐。我们相信，对齐方法提供了一种用于导出嵌入空间对之间的对应关系的便利方法，从而提供了一种用于在源空间和目标空间之间引导模糊标签的方法。这个问题适用于但不限于以下领域。

2.1 Language Translation
2.2 语言翻译

Given a word token in a source language, how can we find a translation of that token in a target language? One such way would be to define pairs of tokens in both the source and target language, creating a labeled dataset of translations. Building such a dataset would require numerous human-hours and thus may not scale well to, or even be feasible in, languages with limited lexical resources. To avoid building these hand-labeled datasets, the task of Bilingual Lexical Induction (BLI) aims to learn mappings from a source to target language in an unsupervised or semi-supervised manner [2]. A critical task in machine translation (MT) systems, BLI has been influenced heavily by embedding techniques, beginning with linear maps from one embedding space to another [3]. By finding structural similarities between two monolingual embedding spaces, such linear maps could generalize to new vocabulary tokens and aid in automated translation. Mappings between languages can also be utilized for transfer learning, where a model using embeddings as input features in one language can be re-adapted to another by simply aligning the feature spaces and re-applying the model. The field of cross-lingual word embedding alignment has rapidly grown, both in the number of evaluation benchmarks and strategies, extending past basic linear maps into the realm of deep learning models [4]. In this survey, we consider word-to-word, word-to-sentence and sentence-to-sentence alignment tasks as solutions to language translation problems.

给定源语言中的单词标记，我们如何找到该标记在目标语言中的翻译？一种方法是在源语言和目标语言中定义成对的标记，创建一个带标签的翻译数据集。构建这样一个数据集将需要大量的人工时间，因此可能无法很好地适应词汇资源有限的语言，甚至在这些语言中不可行。为了避免建立这些手工标注的数据集，双语词汇归纳(BLI)的任务旨在以无监督或半监督的方式学习从源语言到目标语言的映射 2].作为机器翻译(MT)系统中的一项关键任务，BLI 深受嵌入技术的影响，从一个嵌入空间到另一个嵌入空间的线性映射开始 3].通过找到两个单语嵌入空间之间的结构相似性，这种线性映射可以推广到新的词汇标记，并有助于自动翻译。语言之间的映射也可以用于迁移学习，其中在一种语言中使用嵌入作为输入特征的模型可以通过简单地对齐特征空间并重新应用该模型来重新适应另一种语言。跨语言单词嵌入对齐的领域在评估基准和策略的数量上都快速增长，将过去的基本线性映射扩展到深度学习模型的领域[4].在这项调查中，我们认为词到词，词到句子和句子到句子对齐任务作为语言翻译问题的解决方案。

2.3 Knowledge Integration
2.4 知识整合

An ontology is a formal specification of the types of objects and relationships between those objects in a given domain. Ontologies are typically developed by ontology engineers with the goal of providing a controlled vocabulary that can be used and reused to provide exact, specific definitions that may be leveraged by humans and machines alike. Given that ontologies are domain specific, we may encounter cases where we wish to merge ontologies across two connected domains, or cases where two ontologists have independently developed ontologies for the same domain. In such cases, techniques and technologies for highlighting similarities and resolving conflicts are required. A goal of ontology integration is to develop a function that takes two ontologies as inputs and output as one merged

and aligned ontology, matching like entities and relations to reduce duplicity and resolve entity ambiguity [5]. Recently, ontologies have been popularized as ‘knowledge graphs’ through a clever re-branding by Google [6]. Knowledge graphs explicitly frame ontologies using graph data types, allowing for advanced data representation algorithms, such as graph embeddings, to be applied for a variety of tasks, including knowledge base completion [7]. Framing ontology integration as a problem of aligning embeddings of distinct knowledge graphs, researchers have developed techniques for aligning entities in separate graphs for the purpose of forming a joint graph, many of which are surveyed in this paper [8]. In this survey, we consider graph-to-graph alignment as a solution to knowledge integration tasks.
本体是给定领域中对象类型和这些对象之间关系的正式规范。本体通常由本体工程师开发，其目标是提供可被使用和重用的受控词汇表，以提供可被人类和机器等利用的准确、具体的定义。假设本体是领域特定的，我们可能会遇到这样的情况，我们希望合并两个相连领域的本体，或者两个本体学家已经为同一个领域独立开发了本体。在这种情况下，需要突出相似性和解决冲突的技巧和技术。本体集成的一个目标是开发一个功能，该功能将两个本体作为输入和输出作为一个合并和对齐的本体，匹配相似的实体和关系以减少重复并解决实体歧义 5]. 最近，通过谷歌巧妙的品牌重塑，本体已经被推广为“知识图”6]. 知识图使用图形数据类型显式地构建本体，允许高级数据表示算法(如图形嵌入)应用于各种任务，包括知识库完成[7]. 将本体集成框架化为对齐不同知识图的嵌入的问题，研究人员开发了对齐独立图中的实体的技术，以形成联合图，本文对其中的许多技术进行了综述 8]. 在这项调查中，我们认为图到图对齐作为知识整合任务的解决方案。

2.5 Text Understanding and Reasoning
2.6 文本理解和推理

For consumer-facing products like online chat support and information retrieval systems, it is important to have an intuitive interface where a non-technical audience can pose a query and be served information relevant to their needs. Chatbots, question answering (QA) and retrieval systems alike need to be able to interpret user input into a series of intents, understanding the semantic roles of those intents, as well as parse out syntactic language clues such as negation. We call this broad umbrella of applications text understanding and reasoning. Many information reasoning mechanisms are built into knowledge graphs, allowing for the traversals through edges to arrive at facts and inferences. However, access to those systems typically requires technical expertise. In order to leverage the power of knowledge graph semantic reasoners, techniques need to be adapted for understanding, including, but not limited to, slot filling, semantic role labeling and sentence analogies. Harmonizing free text input and understanding with information in a knowledge graph can be seen as an alignment between these two resources, helping to improve the usability and accuracy of these systems. In this survey, we consider word-to-graph and sentence-to-graph alignments as potential solutions to these tasks.
对于面向消费者的产品，如在线聊天支持和信息检索系统，重要的是要有一个直观的界面，让非技术受众可以提出问题，并获得与其需求相关的信息。聊天机器人、问答(QA)和检索系统都需要能够将用户输入解释为一系列意图，理解这些意图的语义角色，以及解析出句法语言线索，如否定。我们称这种广泛的应用为文本理解和推理。许多信息推理机制被构建到知识图中，允许通过边的遍历来得出事实和推论。然而，访问这些系统通常需要专业技术知识。为了利用知识图语义推理机的能力，需要调整技术以进行理解，包括但不限于槽填充、语义角色标记和句子类比。将自由文本输入和理解与知识图中的信息协调起来，可以看作是这两种资源之间的一种结合，有助于提高这些系统的可用性和准确性。在这个调查中，我们认为单词到图形和句子到图形的对齐是这些任务的潜在解决方案。

2.7 Information Extraction

2.8 信息提取

The ability to turn unstructured text data into structured, machine-operable data is a key motivation behind natural language processing tasks such as part-of-speech tagging, named-entity recognition and relation extraction, all of which can be seen as belonging to the task of information extraction. The structured information extracted can be used as inputs to many downstream tasks such as question answering, fact verification and human-computer interactions through chatbots. Here, we focus on the task of relation extraction: the ability to identity two entities and the relation between them from raw text data. The issue of supervised dataset construction is especially detrimental to research in this area; not only are labeled instances hard to collect as human labelers are known to have low precision for the task [9, 10], especially in domains such as biomedicine where subject matter experts are required, but the speed at which new entities and relations may be discussed outpaces the development of such datasets, making supervised models stale and unable to generalize quickly. To combat the problem of data collection, the technique of distant supervision was introduced [11] allowing for entity and relation triples from an ontology or knowledge graph to be used for quick label generation over raw text inputs. In this paradigm, when two entities related by a relation in the knowledge graph appear in a sentence, that sentence is labeled as being representative of that relation. This 'distant supervision' assumption and its relaxations allow for quick bootstrapping of labeled datasets, but it is well known that they also introduce a great deal of noise, as is the case when a sentence mentioning to entities expresses a novel relation, causing a false label, and are equally susceptible to missing labels when the surface forms don't match or are ambiguous [12, 13]. While distant supervision techniques still depend largely on linking the ontology and text corpus via surface forms (i.e. matching on likely string spans or candidate mentions), we anticipate a growing field of alignment between ontology and language embeddings given the increases in alignment techniques used in the two distinct data domains, a main motivating factor for the undertaking of this survey. The task of information extraction serves as our main motivation for undertaking this survey. In this light, we focus our efforts on describing techniques or gaps in current research related to graph-to-sentence alignments as this closely mirrors the task of distant supervision.

将非结构化文本数据转换为结构化的机器可操作数据的能力是自然语言处理任务背后的关键动机，例如词性标注、命名实体识别和关系提取，所有这些都可以被视为属于信息提取的任务。提取的结构化信息可以用作许多下游任务的输入，如问答、事实验证和通过聊天机器人的人机交互。这里，我们关注关系抽取的任务：从原始文本数据中识别两个实体以及它们之间关系的能力。监督数据集构建的问题对该领域的研究尤其不利；不仅被标记的实例难以收集，因为已知人类标记器对于该任务具有低精度[9,10]，特别是在生物医学等需要主题专家的领域，但新实体和关系的讨论速度超过了此类数据集的发展速度，使得监督模型过时，无法快速推广。为了解决数据收集的问题，引入了远程监控技术[11]允许来自本体或知识图的实体和关系三元组用于在原始文本输入上快速生成标签。在这个范例中，当由知识图中的关系相关的两个实体出现在一个句子中时，该句子被标记为代表该关系。这种"远程监督"假设及其放宽允许标记数据集的快速引导，但是众所周知，它们也引入了大量噪声，例如当提及实体的句子表达新的关系时，会导致错误的标记，并且当表面形式不匹配或不明确时，同样容易丢失标记[12,13].虽然远程监督技术仍然很大程度上依赖于通过表面形式链接本体和文本语料库(即，匹配可能的字符串跨度或候选提及)，但鉴于两个不同数据领域中使用的对齐技术的增加，我们预计本体和语言嵌入之间的对齐领域将会增长，这是开展本次调查的主要推动因素。信息提取的任务是我们进行这次调查的主要动机。在这种情况下，我们集中精力描述当前研究中与图形到句子对齐相关的技术或差距，因为这密切反映了远程监督的任务。

2.9 Outline of Survey
2.10 调查大纲

With these motivations in mind, we proceed by first presenting several of the basic methods for generating embedding spaces for language and knowledge graph data in Section 2. We then move on to describe methods for aligning these spaces in Section 3, presenting six situations in which alignments are useful and discussing existing research in these areas, where applicable. We use Section 4 to provide a categorization of approaches to embedding alignment learning. Section 5 discusses benchmark datasets used in this area of research. We conclude in Section 6 with a summary and areas of future research.

考虑到这些动机，我们首先在小节中介绍几种为语言和知识图数据生成嵌入空间的基本方法 2. 然后，我们继续描述对齐这些空间的方法 3，提出六种对齐有用的情况，并讨论这些领域的现有研究。我们使用截面 4 提供嵌入对齐学习方法的分类。部分 5 讨论该研究领域中使用的基准数据集。我们在第一节结束 6 总结和未来的研究领域。

3 Embedding Models
4 嵌入模型

In this section, we outline the basic methods for embedding language and knowledge data into low-dimensional vector spaces.

在本节中，我们概述了将语言和知识数据嵌入低维向量空间的基本方法。

4.1 Word Embedding Models
4.2 单词嵌入模型

Words can be viewed as an atomic unit of natural language. Taking this viewpoint, creating features for machine learning models that involve language can be time consuming. These features typically can include one-hot representations of words or counts of words involved in a sentence or document, both of which suffer from high-dimensionality and sparsity. Finding dense, lower-dimensional representations of words to replace traditional features is the focus of work on word embeddings. The first such modern approach was word2vec, an approach that leverages a shallow neural network to generate hidden state representations. By training such a network, co-occurrences between words are learned to project like-words into the same areas of the low-dimensional space, reflecting their syntactic and semantic properties. Two similar formulations, the continuous bag-of-words model (CBOW) and Skip-gram models were proposed by [14]; here, we focus on the Skip-gram model. We first define the probability of a word w_i given another word w_j as

单词可以被视为自然语言的原子单位。从这个角度来看，为涉及语言的机器学习模型创建特征可能很耗时。这些特征通常可以包括句子或文档中涉及的单词或单词计数的一键表示，这两者都受到高维度和稀疏性的影响。寻找密集、低维的单词表示来代替传统特征是单词嵌入工作的重点。第一个这样的现代方法是 word2vec，这是一种利用浅层神经网络来生成隐藏状态表示的方法。通过训练这样的网络，单词之间的共现被学习来将相似的单词投射到低维空间的相同区域，反映它们的句法和语义属性。[提出了两个类似的公式，连续词袋模型(CBOW)和跳格模型 14];在这里，我们集中讨论跳格模型。我们首先定义一个词 wi 给定另一个词 wj 的概率为

$$p(w_i \mid w_j) = \frac{\exp(u_{w_i}^{\top} v_{w_j})}{\sum_{l=1}^{V} \exp(u_{w_l}^{\top} v_{w_j})}$$

$$p(w_i \mid w_j) = \frac{\exp(u_{w_i}^{\top} v_{w_j})}{\sum_{l=1}^{V} \exp(u_{w_l}^{\top} v_{w_j})}$$

$$p(w \mid w) = \sigma \exp(uwTi \; vwj)$$

where u_w is the trainable vector of input probabilities and v_w the trainable vector of output probabilities for a given word w, and V is the entire vocabulary of the given language domain. Given that the size of V is typically very large, these probabilities are estimated by leveraging negative sampling where noise words are used to generate large contrast

其中 uw 是给定单词 w 的输入概率的可训练向量，vw 是输出概率的可训练向量，V 是给定语言域的整个词汇。假设 V 的大小通常非常大，这些概率是通过利用负采样来估计的，其中噪声字用于产生大的对比度

against potential high probability guesses, eliminating the need to estimate these probabilities over the entire vocabulary for each word. With a sequence of words $w_1, w_2, \ldots, w_n$, the goal of the Skip-gram model is to maximize

针对潜在的高概率猜测，消除了在整个词汇表中为每个单词估计这些概率的需要。具有单词序列 w1，w2，。。。，wn，跳格模型的目标是最大化

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{k} \log p(w_{t+j}$$

对数 p（w $_{t+j}$

$t = 1 \quad j = k$

$|w_t)$ |重量)

where k specifies the window size, i.e. how far to the left and right of the centered word w_t we look when calculating the probability.

其中 k 指定窗口大小，即当计算概率时，我们在中心单词 wt 的左边和右边看多远。

Advances and improvements over this model are plentiful. In [15] the authors develop the GloVe model, making use of a global co-occurrence matrix to address information lost by focusing on small windows during the training of word2vec models. To make the training robust to spelling errors and easier to apply to unseen words at training time, the authors of [16] build FastText, using sub-word embeddings made up of units of characters. Addressing the problem of polysemy, bi-directional neural network models are utilized to capture further context about word usage. Using the Transformer module of [17], word embeddings have matured from *static* representations of those in word2vec, GloVe and FastText, to *contextual* representations used in BERT [18], ELMo [19] and GPT-2 [20]. These models are massive in neural architecture, capturing long distance dependencies between words and interactions of words in multiple contexts. Because of their size, they have the requirement of very large training sets. Due to data and architecture size, these models are typically pre-trained, where initial weights are learned and shared, then re-trained, or fine-tuned, for target tasks or datasets. The availability of these pre-trained models though several APIs has made their use the status quo for natural language processing tasks.

这个模型有很多进步和改进。在[15]作者开发了 GloVe 模型，利用全局共现矩阵来解决在 word2vec 模型的训练期间由于关注小窗口而丢失的信息。为了使训练对拼写错误具有鲁棒性，并且在训练时更容易应用于看不见的单词,[16]构建快速文本，使用由字符单元组成的子词嵌入。为了解决一词多义的问题，双向神经网络模型被用来捕获关于单词使用的进一步的上下文。使用[17]，单词嵌入已经从 word2vec、GloVe 和 FastText 中的静态表示发展到 BERT [18]，ELMo [19]和 GPT-2 [20].这些模型在神经架构中是巨大的，捕捉单词之间的长距离依赖和多个上下文中单词的交互。由于它们的规模，它们需要非常大的训练集。由于数据和体系结构的大小，这些模型通常是预训练的，学习和共享初始权重，然后针对目标任务或数据集重新训练或微调。这些预训练模型的可用性通过几个 API 使得它们的使用成为自然语言处理任务的现状。

4.3 Sentence Embedding Models
4.4 句子嵌入模型

Building on the successes of word embedding models, a logical next step is to use words as atomic units that are composed into sentences. In this shift, moving from a discrete world where words typically represent a handful of semantic units to a continuous representation in a sentence or document, where words can be combined in infinitely many ways, represents a significant challenge. Beginning with sub-word embeddings trained using FastText, an efficient sentence classification model is established in [16] by representing a sentence as the average of its component word representations. Taking the average or sum of static vectors is a common approach to move from word representations to sentence representations, as discussed in [21], and can often beat more advanced models while retaining a level of simplicity. Even though these representations provide successful baselines, they throw out an important element of data in moving from words to sentences: word ordering. To address this, [22] propose utilizing a discrete cosine transformation. By stacking the individual word vectors $w_1, \ldots, w_N$ into a matrix, the discrete cosine transformation can be applied column-wise: for a given column $c_1, \ldots, c_N$, a sequence of coefficients can be calculated as

基于单词嵌入模型的成功，合乎逻辑的下一步是使用单词作为组成句子的原子单位。在这种转变中，从单词通常代表少数语义单位的离散世界转变为句子或文档中的连续表示，单词可以以无限多种方式组合，这是一个重大挑战。从使用 FastText 训练的子词嵌入开始，在[16]通过将句子表示为其组成单词表示的平均值。取静态向量的平均值或总和是从单词表示转移到句子表示的常用方法，如[21]，并且通常可以击败更高级的模型，同时保持一定程度的简单性。尽管这些表示提供了成功的基线，但在从单词到句子的过程中，它们丢掉了一个重要的数据元素:单词排序。为了解决这个问题，[22]建议利用离散余弦变换。通过堆叠各个单词向量 w1，。。。wN 转换成矩阵，离散余弦变换可以逐列应用:对于给定的列 c1，。。。，cN，系数序列可以计算为

and
和

coef [0] =
系数[0] =

网络（Communicating Net的缩写）

普通

通信

$$coef[k] = \frac{1}{N}\sum_{n=0}^{N} c_n \cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right]$$

$$coef = \frac{1}{N}\sum_{n=0}^{N} c_n \cos\left[\frac{\pi}{N}(n+\right.$$

The choice of *k* typically ranges from 0 to 6, where a *k* of zero is essentially the same as vector averaging, while higher orders of *k* account for greater impacts of word sequencing.

k 的选择范围通常从 0 到 6，其中 k 为零基本上与向量平均相同，而 k 的阶数越高，对单词排序的影响越大。

Alternative approaches abandon static word vectors and focus on the sequence of words in the sentence as the starting point. To accommodate sequential data, recursive neural networks (RNNs) dominated the field for a period of time. Recurrent neural network architectures provide an added benefit in that they can theoretically process sequences of variable length up (in practice, this is up to some max length where other sequences are padded with a special token), allowing them to train on corpora with long and short sentences. Another key advantage of RNNs is their ability to share parameters over time where signal from a prior word carries forward to the next word, and so on. This benefit of carrying information forward through the network has a downside of making them hard to train, as gradients need to be

替代方法放弃了静态的单词向量，而将重点放在句子中的单词序列上作为起点。为了适应顺序数据，递归神经网络(RNNs)在一段时间内主导了该领域。递归神经网络架构提供了额外的好处，因为它们理论上可以处理可变长度的序列(在实践中，这可以达到某个最大长度，其中其他序列用特殊标记填充)，允许它们在具有长句子和短句的语料库上进行训练。rnn 的另一个关键优势是能够随着时间的推移共享参数，其中来自前一个字的信号传递到下一个字，依此类推。这种通过网络传递信息的好处有一个缺点，那就是很难训练它们，因为梯度是需要的

propagated backward through time; this has caused them to fall out of favor. One of the first such RNNs trained for sentence encoding was the Skip-Thought model [23]. Rather than use word-context windows, the Skip-Thought model generates an encoding for a center sentence and uses that encoding to predict k sentences to the left and right, where k is again the window size as in the Skip-gram model. To accomplish this, the model leverages an encoder-decoder architecture where each encoder step takes the sequence of words in the sentence and represents them as a hidden state, which is then encoded through the RNN. Decoding then takes place in two steps, one for predicting the next sentence and one for the prior sentence, each of which generates a hidden state through time that can be used to calculate the probabilities of each word in the sequence, with the following objective function

通过时间向后传播；这导致他们失宠。为句子编码训练的第一个这样的 rnn 之一是跳过思维模型[23].Skip-though 模型不是使用单词上下文窗口，而是为中心句子生成编码，并使用该编码来预测左右 k 个句子，其中 k 也是 Skip-gram 模型中的窗口大小。为了实现这一点，该模型利用了编码器-解码器架构，其中每个编码器步骤获取句子中的单词序列，并将它们表示为隐藏状态，然后通过 RNN 进行编码。然后，解码分两步进行，一步用于预测下一句，一步用于预测前一句，每一步生成一个随时间变化的隐藏状态，该隐藏状态可用于计算序列中每个单词的概率，目标函数如下

$$\log P(\text{重量} \quad \sum_{t_t} \log P(w_{i+1}^t \mid w^{<t}, h_i) + \sum_{t_t} \log P(w^t \mid w\langle t, hi\rangle + \sigma \log P(wt$$

$|w \quad , h_i)$

$|w_{,i+1}$ 嗨$)$ $i-$ $<t$ $i-$

An extension of Skip-Thought is the Quick-Thought model [24]. The authors note that the objective function used in Quick-Thought is focused on re-creating the surface forms of each sentence given its dependence on the individual words represented. Specifically, the authors claim "there are numerous ways of expressing an idea in the form of a sentence. The ideal semantic representation is insensitive to the form in which meaning is expressed" [24]. The objective function of Quick-Thought is thus changed to focus only on sentence representations, using a discriminative function to predict a correct center sentence given a window of context sentences. The authors of [25] continue on the quest to capture sentence semantics with the InferSent model. InferSent uses a supervised learning paradigm where sentences in the training set are fed into a three-way classifier, predicting the degree of their similarity (similar, not similar, neutral). Coupled with a bi-directional LSTM model, the InferSent model can be pre-trained on natural language inference (NLI) tasks such as sentence semantic similarity and later used for inference or fine-tuning on other tasks.

快速思维模式是跳跃式思维的延伸 24].作者指出，快速思考中使用的目标函数侧重于重新创建每个句子的表面形式，因为它依赖于所代表的单个单词。具体来说，作者声称"用句子的形式表达一个想法有很多种方式。理想的语义表征对意义表达的形式不敏感 24].因此，快速思考的目标函数被改变为仅关注句子表示，在给定上下文句子窗口的情况下，使用判别函数来预测正确的中心句子。的作者 25]继续探索用推理模型捕捉句子语义。InferSent 使用监督学习范式，将训练集中的句子输入三向分类器，预测它们的相似度(相似、不相似、中性)。与双向 LSTM 模型相结合，推断模型可以在自然语言推断(NLI)任务(如句子语义相似性)上进行预训练，然后用于其他任务的推断或微调。

Similar to the InferSent model, Sentence-BERT uses supervised sentence pairs to learn a similarity function [26]. The input sentence embeddings used for the three-way similarity classifier are generated from a pre-trained BERT model. Contextual models such as BERT provide an encoding of each positional word in an input sentence as their output, thus it is necessary to aggregate these contextualized representations into a single static sentence representation. As with word embedding models, this aggregation can be a sum or average of the representations at a particular layer of the language model, typically the top layer or final layer. An alternative option is to provide the model a special classification token "[CLS]" that has been pre-trained to compress the contextual representations into one layer, which can then be fed to a non-linear unit. Sentence-BERT also adapts the classification task to one of regression where cosine similarity scores are used to score the degree of similarity between sentences. This approach is useful for particular applications, such as semantic search.

与推断模型类似，句子-BERT 使用受监督的句子对来学习相似性函数[26].用于三向相似性分类器的输入句子嵌入是从预先训练的 BERT 模型中生成的。诸如 BERT 之类的上下文模型提供输入句子中每个位置词的编码作为它们的输出，因此有必要将这些上下文化的表示聚合成单个静态句子表示。与单词嵌入模型一样，这种聚集可以是语言模型的特定层(通常是顶层或最终层)的表示的总和或平均值。另一种选择是为模型提供特殊的分类标记"[CLS]"，该标记已经被预先训练以将上下文表示压缩到一个层中，然后该层可以被馈送到非线性单元。Sentence-BERT 还将分类任务调整为回归任务之一，其中余弦相似性分数用于对句子之间的相似程度进行评分。这种方法对于特定的应用是有用的，比如语义搜索。

Continuing on the path of larger, deeper architectures powered by more data, [27] train a Bi-directional LSTM model on a massive scale, multilingual corpus to generate sentence embeddings. Using parallel sentences across 93 input languages, the authors were able to focus on mapping semantically similar sentences to close areas of the embedding space, allowing the model to focus more on meaning and less on syntactic features. Each layer of the LASER model is 512 dimensional, with an output concatenation of both the forward and backward representation generating a final sentence representation of dimension 1024. The model outperforms BERT-like architectures for a variety of tasks including cross-lingual natural language inference, a task focused on detecting sentence similarities.

继续走由更多数据驱动的更大、更深入的架构之路，[27]在大规模、多语言语料库上训练双向 LSTM 模型，以生成句子嵌入。使用跨 93 种输入语言的平行句子，作者能够专注于将语义相似的句子映射到嵌入空间的封闭区域，从而允许模型更多地关注意义，而不是句法特征。激光器模型的每一层都是 512 维的，正向和反向表示的输出连接生成了 1024 维的最终句子表示。该模型在各种任务上优于 BERT-like 架构，包括跨语言自然语言推理，这是一项专注于检测句子相似性的任务。

4.5 Knowledge Graph Embedding Models
4.6 知识图嵌入模型

Building on the success of embedding-based methods in natural language processing, these techniques have spilled into the domain of knowledge graphs. Their main motivating uses are in the task of statistical representation learning where the larger graph is compressed into a low-dimensional representation that can be used by reasoning systems,

and knowledge base completion (KBC), where embeddings of existing facts can be utilized to predict new relationships between entities in the graph. Approaches in this area can be classified into three main categories: translation-based models, semantic-matching models and graph-structure models. Our aim is to introduce models leveraged in the alignment literature; for more comprehensive introductions see [7] and [28].

基于嵌入的方法在自然语言处理中的成功，这些技术已经扩展到知识图的领域。它们的主要激励用途是在统计表示学习任务中，其中较大的图形被压缩成推理系统可以使用的低维表示，以及知识库完成(KBC)，其中现有事实的嵌入可以被用来预测图形中实体之间的新关系。这一领域的方法可以分为三大类:基于翻译的模型、语义匹配模型和图结构模型。我们的目的是介绍比对文献中利用的模型；有关更全面的介绍，请参见[7] 和[28].

4.6.1 Translation-based Methods
4.6.2 基于翻译的方法

Let $G = (E, R)$ be a knowledge graph consisting of a set of entities E and relations R, each element of which may have an entity or relation type. From this graph, we can construct the set of known facts, represented as triples (h, r, t) with $h, t \in E$ and $r \in R$. The intuition behind translation-based models is we wish to have low-dimensional, dense representations of h, r, t such that $h + r \approx t$. Model choices then depend on which space or spaces the entities and

设 G = (E，R)是由一组实体 E 和关系 R 组成的知识图，其中每个元素可能有一个实体或关系类型。从该图中，我们可以构建已知事实的集合，用三元组(h，r，t)和 h，t ∈ E 和 r ∈ R 表示。基于翻译的模型背后的直觉是，我们希望具有 h，r，t 的低维、密集表示，使得 h + r ≈ t

relations are embedded in as well as the scoring function used to help the model learn to differentiate between true triples from the graph and noise triples that do not reflect real-world facts. TransE [29] is the simplest of these models. It embeds both the entities and relations in the same low-dimensional vector space and uses a simple distance function defined by

关系以及用于帮助模型学习区分来自图的真实三元组和不反映真实世界事实的噪音三元组的评分函数被嵌入。TransE [29]是这些模型中最简单的。它将实体和关系都嵌入到同一个低维向量空间中，并使用由定义的简单距离函数

$$f_r(h, t) = -\|h + r - t\|_{1/2}$$

$$fr(h, t) \nparallel h+r\ t1\|/2$$

While this model is simple, it struggles to properly encode one to many triples, where a single relation may hold between a head entity and several tail entities. Resolutions to this are addressed in TransH [30], where each relation is assigned its own hyperplane. Similarly, TransR [31] lets each relation have its own distinct embedding space ad thereby greatly expands the parameter space of the model and increases the capability of learning relation-specific translations. An entire family of translation-based models exists, each adding various complexities and constraints to the embedding spaces with novel loss functions used to best recover the relations described in the original graph.

虽然这个模型很简单，但它很难正确地编码一个到多个三元组，其中一个头部实体和几个尾部实体之间可能只有一个关系。对此的解决方案在 TransH30]，其中每个关系被分配给它自己的超平面。同样，TransR [31]使得每个关系具有其自己独特的嵌入空间，从而极大地扩展了模型的参数空间，并增加了学习特定于关系的翻译的能力。存在一整个系列的基于翻译的模型，每一个都用新颖的损失函数向嵌入空间增加了各种复杂性和约束，所述损失函数用于最好地恢复原始图中描述的关系。

4.6.3 Semantic-matching Models
4.6.4 语义匹配模型

While the translational assumption $h + r \approx t$ gives good geometric intuition as to the types of relations learned during model training, it is prohibitive for a wide class of relations, including those with anti-symmetric or complex properties. Semantic-matching models deviate away from the distance-based assumption and focus on using similarity-based scoring functions to recover facts from the low-dimensional representations of entities and relations. Rather than relying on norms and translations, these methods leverage dot-product like scoring functions to measure angles between low-dimensional representations, sometimes referred to as 'semantic energy' functions. The simplest of such models is RESCAL [32], which relies on a tensor representation of the underlying knowledge graph X, where each entry of the tensor $X_{ijk} = 1$ if the fact is represented in the knowledge graph, else-wise zero. This tensor can then be factorized into latent components,

虽然平移假设 h + r t 对于在模型训练期间学习的关系类型给出了良好的几何直觉，但是它对于广泛的关系类别是禁止的，包括具有反对称或复杂属性的关系。语义匹配模型偏离了基于距离的假设，并专注于使用基于相似性的评分函数来从实体和关系的低维表示中恢复事实。这些方法不依赖于规范和翻译，而是利用点积之类的评分函数来测量低维表示之间的角度，有时被称为“语义能量”函数。这类模型中最简单的是 RESCAL[32]，其依赖于基础知识图 X 的张量表示，其中如果事实在知识图中被表示，则张量的每个条目 Xijk = 1，否则为零。这个张量然后可以分解成潜在分量，

$$X_k \approx AR_kA^T \text{ for } k = 1, \ldots, r$$

$$对于 k = 1，Xk \approx ARkAT。。。。，r$$

where R_k is a matrix of dimension $r \times r$ representing interactions between each corresponding component and A contains the r dimensional representations of the entities. Thus for each (h, t) pair, we can compute the likelihood they participate in the k-th relation as

其中 Rk 是 r 维矩阵，r 表示每个相应组件之间的交互，A 包含实体的 r 维表示。因此，对于每个(h，t)对，我们可以如下计算它们参与第 k 个关系的可能性

$$f_k(h, t) = h^{\top} R_k t$$

$$fk(h, t) = hTRkt$$

Contrasted with translation-based models, RESCAL takes advantage of vector products while capturing interactions between elements of each entity and all relations. In a simplification, DistMult [33] requires each R_k to be diagonal, reducing the parameters of the model while sacrificing some of its representational capacity. This reduction in capacity is especially felt when modeling anti-symmetric relations as interactions in these diagonal matrices have no notion of directionality. To circumvent this issue, the ComplEx [34] model allows for the low-dimensional representations to live in the complex space $\mathbb{C}$. The scoring function used by the ComplEx model is defined as

与基于翻译的模型相比，RESCAL 在捕捉每个实体的元素和所有关系之间的交互时利用了矢量积。在简化中，DistMult [33]要求每个 Rk 都是对角的，减少了模型的参数，同时牺牲了一些表示能力。当对反对称关系建模时，容量的减少尤其明显，因为在这些对角矩阵中的相互作用没有方向性的概念。为了避免这个问题，复杂[34]模型允许低维表示存在于复杂空间 c 中。复杂模型使用的评分函数被定义为

$$f_k(h, t) = \Box(|h, w_k, t|) \text{ where } w_k \in \mathbb{C}^r.$$

$$fk(h, t) = (|h, wk, t|) 其中 wk \in Cr。$$

By allowing the representations to be complex-valued, the model can handle the asymmetries of many relations present in knowledge graphs, yet score the likelihoods of facts existing using only the real-valued vectors. The work of [35] takes this one step further, defining the ConvE model where entities interact through the convolution operator. This introduces additional non-linearities through which the model can increase the capacity for learning complicated relational structures.

通过允许表示为复值，该模型可以处理知识图中存在的许多关系的不对称性，还可以仅使用实值向量对存在的事实的可能性进行评分。的工作 35]更进一步，定义了 ConvE 模型，其中实体通过卷积运算符进行交互。这引入了额外的非线性，通过该非线性，模型可以增加学习复杂关系结构的能力。

4.6.5 Graph-structure Models
4.6.6 图形结构模型

Given that the entities and relations between them are modeled as a graph with vertices and edges, we can take advantage of the underlying graph structure to aide in creating low-dimensional representations. Graph representation learning has been a trending topic over the past few years with many advances in creating representations of graphs that can be used in machine learning models. We refer the reader to [36] for a complete introduction.

假设实体和它们之间的关系被建模为具有顶点和边的图，我们可以利用底层的图结构来帮助创建低维表示。在过去几年中，随着在创建可用于机器学习模型的图形表示方面的许多进步，图形表示学习已经成为一个趋势性的主题。我们请读者参考[36] 完整的介绍。

For knowledge graph embedding, path traversal techniques have been applied to learn additional facts about the relations between multiple entities in a graph, rather than just the one-hop paths learned in translation-based and

对于知识图嵌入，已经应用路径遍历技术来学习关于图中多个实体之间关系的附加事实，而不仅仅是在基于翻译的和

semantic-matching models. By taking a walk on the graph we can learn more about the neighborhood structures of each entity, using that information to learn better dense representations. Using paths on the graph, the PTransE approach extends the traditional TransE method to capture structural information [37]. Given two entities h and t and a path $p = r_1 \rightarrow r_2 \rightarrow \ldots \rightarrow r_k$, where each r_i is a relational embedding, the authors of PTransE propose three ways of aggregating all relation vectors involved to a path vector. These include addition: $p = r_1 + \ldots + r_k$, multiplication: $p = r_1, \ldots, r_k$ and application of a RNN: $c_i = f(W[c_{i-1} : r_i])$ where f is a non-linearity and [:] represents vector concatenation.

语义匹配模型。通过在图上走一走，我们可以更多地了解每个实体的邻域结构，使用该信息来学习更好的密集表示。使用图上的路径，PTransE 方法扩展了传统的 TransE 方法来捕获结构信息[37]. 给定两个实体 h 和 t 以及一条路径 p = r1 r2。。。rk，其中每个 ri 是一个关系嵌入，PTransE 的作者提出了三种将所有涉及的关系向量聚合成一个路径向量的方法。这些包括加法:p = r1 +。。。+ rk，乘法:p = r1。。。rk 和 RNN 的应用:ci = f (W [ci 1 : ri])，其中 f 是非线性的，而[:]表示向量级联。

In a similar approach, [38] create entirely new triples from paths in the knowledge graph. If h and t are connected by the path $p = r_1 \rightarrow \ldots \rightarrow r_k$, they add a new triple (h, p, t) to the set of known triples used to train the knowledge graph. The triples can now be recovered using the translation-based loss of TransE

以类似的方式，[38]从知识图中的路径创建全新的三元组。如果 h 和 t 由路径 p = r1 连接。。。rk，他们将一个新的三元组 h，p，t 添加到用于训练知识图的已知三元组集合中。现在可以使用基于翻译的 TransE 丢失来恢复三联体

$$f_k(h, t) = -\|h + (r_1 + \ldots + r_k) - t\|_{1/2}$$

fk(h， t)=-||h+(R1+。。。+rk) t||1/2

or used in the RESCAL context through multiplication of the relevant slices of the factorized matrix R_k

或者通过因子分解矩阵 Rk 的相关切片的乘法在重缩放上下文中使用

$$f_k(h, t) = h^{\mathsf{T}}(M_1 \cdot \ldots \cdot M_k)t$$

fk(h， t)= M1。。。Mk) t

where each $M_i \in R_k$.

其中每个 Mi ∈ Rk。

More recently, attention has turned to using graph convolutional networks [39] (GCNs) for knowledge graph embeddings. These techniques are called convolutional as they use neighborhood features of each node, similar to how convolutional operators look at borders of each pixel in computer vision models. By representing the knowledge graph G by its adjacency matrix A and let X be a matrix of representations (features) of the entities in the graph. The convolutions, defined layer by layer, can be represented as

最近，注意力已经转向使用图卷积网络[39] (GCNs)用于知识图嵌入。这些技术被称为卷积，因为它们使用每个节点的邻域特征，类似于卷积运算符在计算机视觉模型中查看每个像素边界的方式。通过用邻接矩阵 A 表示知识图 G，并且让 X 是图中实体的表示(特征)的矩阵。逐层定义的卷积可以表示为

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)})$$

h(1+1)=σ (D\\—1/2a \\d\\—1/2H(1)W(1))

for the $l+1$th layer, where $\tilde{A} = A + I_N$ and I_N is the identity matrix, $\tilde{D}_{ii} = \Sigma_j \tilde{A}_{ij}$ and $W^{(l)}$ is the weight matrix for each layer. Here, $H^{(0)} = X$, meaning the process begins by considering individual nodes and expands to represent their neighborhoods up to the number of layers in the network.

对于第 l+1 层，其中 A \= A+IN，IN 是单位矩阵，d \ii =σ ja \ij，W (l)是每层的权重矩阵。在这里，H(0) = X，意味着该过程从考虑单个节点开始，并扩展到表示它们的邻域，直到网络中的层数。

Translation-based methods, semantic-matching models and graph-structure models have all been used to embed individual knowledge graphs as well as aide in entity alignment between embedded graphs, as described in Section 3.4.

基于翻译的方法、语义匹配模型和图结构模型都被用于嵌入单个知识图，以及帮助嵌入图之间的实体对齐，

如第节所述 3.4.

5 Alignment Approaches

6 对齐方法

Abstractly, learning a mapping function between two vector spaces is a well studied problem. Let D_1 and D_2 be two datasets, originating from either similar (as is the case for two language corpora from different languages) or different (as is the case for a set of images and a language corpus) domains. Let the functions $f_1 : D_1 \rightarrow \mathbb{R}^n$ and $f_2 : D_2 \rightarrow \mathbb{R}^m$ represent two mappings from the original datasets to their respective real-valued embedding spaces. Typically, n and m are of much lower dimension than the original cardinalities of D_1 and D_2, and therefore f_1 and f_2 can be thought of as techniques to compress the original datasets whilst maintaining their defining geometric characteristics, including a notion of 'semantic similarity'. These similarities are measured in the lower-dimensional vector spaces through techniques such as, but not limited to, Euclidean distance or cosine similarity.

理论上，学习两个向量空间之间的映射函数是一个研究得很好的问题。假设 1 和 2 是两个数据集，源自相似的(如来自不同语言的两个语言语料库的情况)或不同的(如一组图像和一个语言语料库的情况)领域。让函数 f1 : 1 Rn 和 f2 : 2 Rm 表示从原始数据集到它们各自的实值嵌入空间的两个映射。通常，n 和 m 的维数比原始基数 1 和 2 低得多，因此 f 1 和 f2 可以被认为是压缩原始数据集的技术，同时保持它们定义的几何特征，包括“语义相似性”的概念。这些相似性是通过诸如但不限于欧几里德距离或余弦相似性的技术在低维向量空间中测量的。

Let us assume that these 'semantic similarities' are preserved by the functions f_1 and f_2. If there exists a correspondence between elements $x \in D_1$ and $y \in D_2$, then the problem of aligning their respective embedding spaces seeks to find a map $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $A(f_1(x)) \approx f_2(y)$.

让我们假设函数 f1 和 f2 保留了这些“语义相似性”。如果在元素 x ∈ D1 和 y ∈ D2 之间存在对应，那么对齐它们各自的嵌入空间的问题寻求找到映射 A : Rn → Rm，使得 A(f1(x)) ≈ f2(y)。

More generally, these methods seek to detect and exploit *invariances* between pairs of low-dimensional embedding spaces. The degree to which these invariances can be captured dictates how much training data is required to learn a reliable alignment model. In the case where the underlying geometric structures of both embedding spaces are perfectly invariant, up to a rotation of the space, simple maps may be learned in a highly unsupervised way. However, on the flip-side of the coin, methods which do not generate well structured embedding spaces may require more training data in order to learn alignments. Critically, the problem of learning an alignment map A is also tied to the choice of good embedding functions f_1 and f_2, and careful coordination between all three choices is required for finding an optimal solution.

更一般地，这些方法寻求检测和利用低维嵌入空间对之间的不变性。可以捕获这些不变性的程度决定了需要多少训练数据来学习可靠的对齐模型。在两个嵌入空间的基本几何结构完全不变的情况下，直到空间旋转，简单的映射可以以高度无监督的方式学习。然而，在硬币的另一面，不产生良好结构的嵌入空间的方法可能需要更多的训练数据来学习比对。至关重要的是，学习比对图 A 的问题也与选择好的嵌入函数 f1 和 f2 相关联，并且为了找到最优解，需要在所有三个选择之间进行仔细的协调。

Table 1: Keyword Labels for Research Classification

表 1:研究分类的关键词标签

Label	Keywords
Knowledge Graph	node, knowledge graph, network, ontology, knowledge base
Sentence	sentence, phrase, cross-lingual,
Word	multilingual word, token, cross-lingual, multilingual
标签	关键词
知识图谱	节点，知识图，网络，本体，知识库
句子	句子，短语，跨语言，多语言
单词	单词，令牌，跨语言，多语言

6.1 Research Landscape

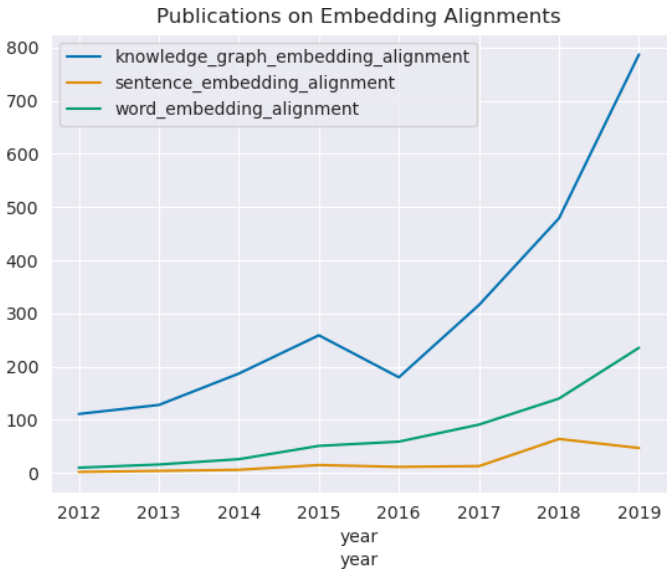
6.2 研究前景

As mentioned in our motivating section, we believe that embedding alignments are critical to the task of information extraction, particularly in mapping unstructured text to structured knowledge graphs. However, our hypothesis is that these techniques have yet to be fully explored in the research community, especially in learning alignments between sentence embeddings and graph embeddings. To understand the landscape of research in alignments, we undertook a search of three research repositories: ArXiv, DBLP and IEEE. Our search dates ranged from January 1, 2012 (picked to cover a period of a year before the publication of the word2vec paper) through the end of 2019 (to avoid a partial year of 2020 at the time of publication). For each repository, we conducted a keyword search for 'embedding alignment'. We then further sub-divided the results into four categories as they pertain to embeddings: knowledge graph, sentence, word or not applicable. These categories are determined by a simple count of keyword matches in the paper's abstract, as outlined in 1, with ties being assigned to both categories.

正如在我们的激励部分提到的，我们认为嵌入比对对于信息提取的任务是至关重要的，特别是在将非结构化文本映射到结构化知识图的过程中。然而，我们的假设是，这些技术尚未在研究社区中得到充分探索，特别是在句子嵌入和图形嵌入之间的学习对齐方面。为了了解比对研究的前景，我们对三个研究资源库进行了搜索:ArXiv、DBLP 和 IEEE。我们的搜索日期范围从 2012 年 1 月 1 日(选择覆盖 word2vec 论文发表前一年的时间)到 2019 年底(以避免发表时 2020 年的部分年份)。对于每个存储库，我们进行了“嵌入比对”的关键字搜索。然后，我们将结果进一步细分为四个类别，因为它们与嵌入相关:知识图、句子、单词或不适用。这些类别是由论文摘要中简单的关键字匹配数决定的，如 1，两个类别都有联系。

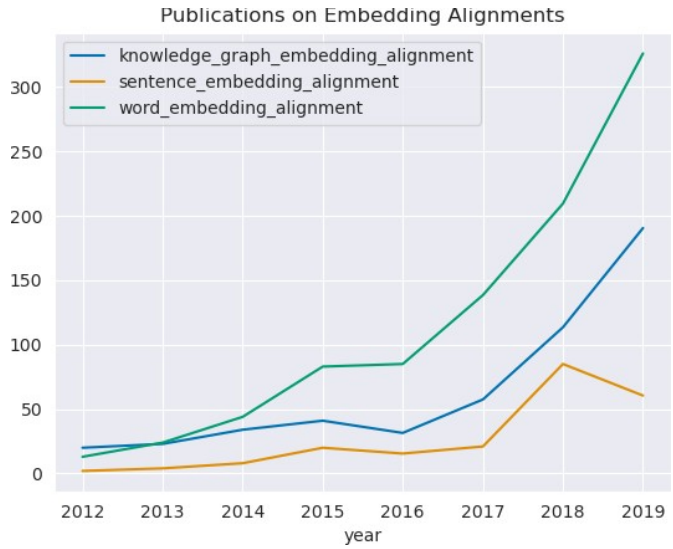
We then measured the trend over time for mentions of word embedding alignments, sentence embedding alignments and knowledge graph embedding alignments. The results are shown in the following figure.

然后，我们测量了单词嵌入比对、句子嵌入比对和知识图嵌入比对的提及随时间的趋势。结果如下图所示。



We note that the inclusion of the ‘network’ keyword artificially inflates the count of publications classified as aligning knowledge graph embeddings. Many of these papers deal with embeddings of social networks, and while those networks could be considered knowledge graphs they do not fit the definition of a knowledge graph used herein. The same trend plot with network related papers removed can be seen as follows.

我们注意到，包括“网络”关键字人为地膨胀了分类为对齐知识图嵌入的出版物的数量。这些论文中的许多涉及社会网络的嵌入，虽然这些网络可以被认为是知识图，但它们不符合这里使用的知识图的定义。移除网络相关论文后的相同趋势图如下。



Per the above trends, we see that there has been a continued rise in the application of both knowledge graph and word embedding technologies. Mentions of sentence embeddings are dwarfed by the other two categories, with some of that trend explained by limitations in 3.2.2. Given the popularity of embedding techniques in the machine learning community, we believe growth in this research area will continue. We also hypothesize that alignments between other objects, i.e. tokens to graphs, will become an increasingly important field for knowledge and data integration. We proceed by enumerating the potential applications of embedding space alignments.

根据上述趋势，我们看到知识图和单词嵌入技术的应用持续增长。与其他两个类别相比，对句子嵌入的提及相形见绌，其中一些趋势可以用 3.2.2. 鉴于嵌入技术在机器学习社区的流行，我们相信这一研究领域的增长将会继续。我们还假设，其他对象之间的比对，即图的标记，将成为知识和数据集成的一个越来越重要的领域。我们通过列举嵌入空间排列的潜在应用来继续。

6.3 Alignment Use Case Enumeration
6.4 对齐用例枚举

Given the focus of this survey on the domains of language and knowledge graphs, we outline six situations in which embedding space alignments could occur. In these cases, we assume that the direction of the learned alignment mapping is irrelevant, i.e. we could easily reverse the source and target spaces and learn an alignment map in the reverse direction.

鉴于本次调查的重点是语言和知识图领域，我们概述了嵌入空间对齐可能发生的六种情况。在这些情况下，我们假设所学习的对齐映射的方向是不相关的，即，我们可以容易地颠倒源和目标空间，并且在相反的方向上学习对齐映射。

6.4.1 Word-to-Word Alignment
6.4.2 词对词对齐

The ultimate goal of a word-to-word alignment model is to be able to input the embedding of a token in a source language and receive as output the embedding of a semantically or syntactically similar token in the target language. As first noted in [3], word embedding models trained on distinct languages exhibited similar geometric patterns and behaviors. This observation led the authors to hypothesize that word embedding spaces could be transformed from one to another through simple linear operations, such as translation and rotation. The first attempts and models in this area took advantage of large, parallel vocabularies, where pairs of words were used to learn mapping matrices from one space to another. While learning the transformation matrix may have a closed-form solution and could

be directly solved through linear algebraic methods, in practice, the weights are learned through stochastic gradient descent. We survey the common supervised learning model types in Section 4.1.1. Given the relative success and ease of implementation of these models when parallel data is available, researchers began to ask how limited that parallel set could be. Restricting to the top 5,000 most common words, restricting the parts of speech, or even relying only on numerals have been popular approaches into reducing the level of supervision needed to learn strong translation models [40]. Hybrid approaches use a form of semi-supervised learning, beginning from small seed lexicons and iteratively adding words as confidence in their direct translation builds. We introduce these semi-supervised methods in Section 4.1.2. Moving past semi-supervised methods, approaches to learning mappings between embedding spaces in a completely unsupervised way. These methods rely on the geometric structures of the underlying spaces as a proxy for parallel data, either relying on embedding similarity distributions [41], adversarial learning [42] or metric recoveries via optimal transport [43]. We cover these methods in Section 4.1.3.

词到词对齐模型的最终目标是能够输入源语言中的标记嵌入，并接收目标语言中语义或句法上相似的标记嵌入作为输出。正如在[3]，在不同语言上训练的单词嵌入模型表现出相似的几何模式和行为。这一观察导致作者假设单词嵌入空间可以通过简单的线性操作，如平移和旋转，从一个转换到另一个。这一领域的第一次尝试和模型利用了大量的并行词汇，其中成对的单词用于学习从一个空间到另一个空间的映射矩阵。虽然学习变换矩阵可以具有封闭形式的解，并且可以通过线性代数方法直接求解，但是在实践中，通过随机梯度下降来学习权重。我们将在第 3 节中介绍常见的监督学习模型类型 4.1.1. 考虑到当并行数据可用时这些模型的相对成功和容易实现，研究人员开始询问并行集可以有多有限。限制到最常见的前 5000 个单词，限制词性，或者甚至仅依赖数字，已经成为降低学习强翻译模型所需的监督级别的流行方法[40].混合方法使用一种半监督学习的形式，从小种子词典开始，并在直接翻译中迭代添加单词。我们将在第 2 节介绍这些半监督方法 4.1.2. 超越半监督方法，以完全无监督的方式学习嵌入空间之间的映射的方法。这些方法依赖于底层空间的几何结构作为并行数据的代理，或者依赖于嵌入的相似性分布[41]，对抗性学习[42]或通过最佳传输进行度量恢复[43]. 我们将在第 5 节中介绍这些方法 4.1.3.

6.4.3 Sentence-to-Sentence Alignment
6.4.4 句子间对齐

Sentence to sentence alignment often serves as an entry point to machine translation applications. Given a parallel corpus of sentences in two languages, the goal is to learn a mapping function f that converts a low-dimensional

句子到句子的对齐通常是机器翻译应用的切入点。给定两种语言的句子的平行语料库，目标是学习映射函数 f

representation of sentence $s_1 \in L_1$ to a close (in terms of vector space proximity) representation $t_1 \in L_2$. This map can then generalize for future translations such that $f(s_2) \approx t_2$ where $s_2 \in L_1$, $t_2 \in L_2$. This approach is limited due to two factors. First, the availability of such parallel corpora is limited. Most research in this area either relies on the Europarl dataset [44] or translations of the Bible. Neither of these resources represents enough diversity in language to scale up to production-level systems, but they do allow for ideas to be tested experimentally. The second limitation comes from the composition of semantic units (i.e. individual word tokens) to higher order representations in sentences. Word order plays a role in the structure of languages, thus simple mapping models have been replaced with those that model the sequences of tokens, such as the seq2seq model [45].

句子 s1 1 的表示接近(根据向量空间接近度)表示 t1 2。然后，这个映射可以推广到将来的翻译，例如 f (s2) t2，s2 L1，t2 L2。这种方法受到两个因素的限制。首先，这种平行语料库的可用性是有限的。该领域的大多数研究要么依赖于 Europarl 数据集[44]或《圣经》的译本。这些资源都没有表现出足够的语言多样性来扩展到生产级别的系统，但是它们允许想法被实验性地测试。第二个限制来自句子中语义单元(即单个单词标记)到高阶表示的组合。词序在语言结构中起着重要作用，因此简单的映射模型已经被那些对标记序列建模的模型所取代，比如 seq2seq 模型[45].

6.4.5 Word-to-Sentence Alignment
6.4.6 词-句对齐

Given the availability of technologies for word-to-word and sentence-to-sentence alignment, there has been little need for additional research in word-to-sentence alignment. In the case of a set of sentences being mapped to a finite set of words, this is typically handled as a supervised classification problem where the finite set of words is one-hot encoded to represent a target variable. We direct the reader to [16] for efficient approaches to this type of supervised classification problem.

鉴于单词到单词和句子到句子对齐技术的可用性，几乎不需要在单词到句子对齐方面进行额外的研究。在一组句子被映射到一组有限的单词的情况下，这通常被作为监督分类问题来处理，其中该有限的单词组被一次性编码以表示目标变量。我们建议读者去[16]以获得解决这类监督分类问题的有效方法。

6.5 Sentence-to-Sentence Alignment
6.6 句子间对齐

While the goal of word-to-word alignment is to map tokens for direct translation, these tokens often can express multiple senses and thereby exhibit polysemy. This creates issues in direct, one-to-one mappings due to the fact that the training data can contain a particular token in the source space with multiple translations in the target space, leading to conflicting information during training and at inference. Rather than focusing on tokens as the atomic unit, tokens in a given context, either through phrases or complete sentences, carry more information that can be leveraged for better alignment. The research area of sentence-to-sentence alignment relies on parallel documents, typically found in resources such as translations of the European Parliament proceedings or the Bible. While alignment of parallel word tokens is a rich research field, there has been less focus on alignment techniques of full sentences; research in this area typically falls under the umbrella of machine translation where more complex sequence-to-sequence neural models are favored. However, many of the same techniques for aligning word embeddings can be leveraged for aligning sentences provided we can generate representative sentence embeddings. We cover a handful of sentence embedding methods in 2.2 and discuss their alignment in section 4.3.

虽然词到词对齐的目标是映射用于直接翻译的标记，但是这些标记通常可以表达多个意思，从而表现出多义性。这在直接的一对一映射中产生了问题，因为训练数据可以包含源空间中的特定标记，而在目标空间中有多个翻译，从而导致训练和推断过程中的信息冲突。给定上下文中的标记，无论是通过短语还是完整的句子，都携带了更多的信息，可以用于更好的对齐，而不是将标记作为原子单位。句子到句子对齐的研究领域依赖于平行文档，通常可以在诸如欧洲议会议事录或圣经的翻译等资源中找到。虽然平行单词标记的对齐是一个丰富的研究领域，但对完整句子的对齐技术关注较少；该领域的研究通常属于机器翻译的范畴，其中更复杂的序列间神经模型更受青睐。然而，如果我们能够生成有代表性的句子嵌入，许多用于对齐单词嵌入的相同技术也可以用于对齐句子。我们在中介绍了一些句子嵌入方法 2.2 并在第节中讨论它们的对齐 4.3.

6.7 Knowledge-to-Knowledge Alignment
6.8 知识到知识的匹配

Knowledge graphs have seen a great deal of interest and hype in recent years as their applications to artificial intelligence and machine learning have come to be seen as the onset of a 'third wave' contributing to semantically grounded and explainable AI. They also serve as the backbone to the Semantic Web, a set of standards for defining and linking data and meta-data in a machine-readable and human-interpretable way. While large corporations like Google and LinkedIn have massive knowledge graphs at their disposal, smaller, more tailored graphs exist for specific purposes such as SnoMed for medical clinical terminology and FIBO for financial industry concepts. Given the specificity of some of these smaller graphs, we may wish to weave several of them together for a particular application, including data exchange protocols and data integration tasks. We may also want to merge knowledge graphs covering similar, yet independently defined, concepts, or graphs defining the same subject matter across languages. This task is referred to in the literature as *entity alignment*: the process of identifying nodes in each graph that are referencing the same semantic concept and either forming relations between them or compressing them into a single representation. Work in this area originated in the task of ontology alignment [46], which aimed to use heuristics, string matching and natural language processing techniques to map source and target nodes. As in the other application areas in this survey, the push to deep, representational learning has invaded the space of ontology alignment as well, typically couched under the banner of entity embedding alignment. The task at hand is to create low-dimensional representations of the source and target knowledge graphs and use only these embeddings to automatically discover alignments. As in word-to-word alignments, methods range from directly supervised methods where parallel entities between graphs are used to learn mappings, to fully unsupervised methods where inferences are made to align entities based on the structure of their neighborhoods in the graph. We cover these techniques in Section4.2

近年来，随着知识图在人工智能和机器学习中的应用被视为“第三次浪潮”的开始，知识图受到了极大的关注和炒作，这有助于语义基础和可解释的人工智能。它们也是语义网的主干，语义网是一套以机器可读和人类可理解的方式定义和链接数据和元数据的标准。虽然像谷歌和 LinkedIn 这样的大公司拥有大量的知识图表，但更小、更定制的图表是为了特定目的而存在的，如医疗临床术语的 SnoMed 和金融行业概念的 FIBO。考虑到这些小图的特殊性，我们可能希望将它们编织在一起用于特定的应用，包括数据交换协议和数据集成任务。我们可能还想合并涵盖相似但独立定义的概念的知识图，或者跨语言定义相同主题的知识图。这项任务在文献中被称为实体对齐:识别每个图中引用相同语义概念的节点并在它们之间形成关系或将它们压缩成单个表示的过程。这个领域的工作起源于本体对齐的任务[46]，旨在使用启发式、字符串匹配和自然语言处理技术来映射源节点和目标节点。正如本次调查中的其他应用领域一样，对深度、代表性学习的推动也侵入了本体对齐的空间，通常打着实体嵌入对齐的旗号。手头的任务是创建源和目标知识图的低维表示，并仅使用这些嵌入来自动发现比对。如同在词到词的对齐中一样，方法的范围从直接监督的方法(其中使用图形之间的平行实体来学习映射)到完全非监督的方法(其中进行推断以基于图形中实体的邻域结构来对齐实体)。我们将在中介绍这些技术 Section4.2

6.8.1 Word-to-Knowledge Alignment
6.8.2 单词-知识对齐

In the previously explored cases, embeddings of source and target data from similar domains were aligned. In these cases, strings are mapped to strings and graph entities to other graph entities. This section deviates from those proceeding by considering alignments from strings to knowledge graph entities. For a given token $x \in D_1$, we wish to identify a

在前面探索的案例中，来自相似域的源数据和目标数据的嵌入是一致的。在这些情况下，字符串被映射到字符串，图形实体被映射到其他图形实体。本节通过考虑从字符串到知识图实体的比对，偏离了前面的内容。对于给定的令牌 $x \in D1$，我们希望确定一个

corresponding entity $y \in$ if such an entity exists. One such way of finding these correspondences is to find a map between the token embedding $f_1(x)$ and the entity embedding $f_2(y)$. Given that the target domain (a knowledge graph) is constructed to reflect facts about real world entities and the relations between them, we expect to find those same facts and entities referred to in the source space (language), although with much lower precision and exactness in their statements. While inherent noise present in human language makes learning such an alignment challenging, success in this area can assist with knowledge driven entity extraction and named entity recognition.

对应的实体 y 2，如果这样的实体存在。找到这些对应的一种这样的方式是找到嵌入 f1(x)的令牌和嵌入 f2(y)的实体之间的映射。假定目标域(知识图)被构造来反映关于真实世界实体的事实和它们之间的关系，我们期望找到在源空间(语言)中引用的那些相同的事实和实体，尽管在它们的陈述中具有低得多的精确度和准确性。虽然人类语言中存在的固有噪声使得学习这样的对齐具有挑战性，但是在该领域中的成功可以帮助知识驱动的实体提取和命名实体识别。

6.8.3 Sentence-to-Knowledge Alignment
6.8.4 句子-知识对齐

Our main motivation in this line of research is the alignment of sentences to knowledge graphs. The interest in this problem is two-fold. Firstly, if we are able to align embeddings of triples (h, r, t) from the knowledge graph G to sentence embeddings s in a given corpora, these alignments can be used to detect the expression of relationships r in the sentences, aiding in the task of relation extraction. Secondly, in the opposite direction, if we can align sentences to triples, we can use this technique to assist in the detection of new triples to be added to the knowledge graph from text data, aiding in the automated expansion of a knowledge graph. These two problem domains can be viewed as complementary techniques for converting unstructured data in text documents to structured data in a knowledge graph. Having data in a structured format not only makes it easier for human verification, as in the case of automated fact checking, but also allows for insights into how other machine learning models, such as document classification, are leveraging unstructured data, providing an avenue for explainable AI and model governance.

我们在这方面研究的主要动机是将句子与知识图对齐。对这个问题的兴趣是双重的。首先，如果我们能够将来自知识图 G 的三元组 h，r，t 的嵌入与给定语料库中的句子嵌入 s(对齐)，这些对齐可以用于检测句子中关系 r 的表达，有助于关系提取的任务。其次，在相反的方向上，如果我们可以将句子与三元组对齐，我们可以使用这种技术来帮助检测要从文本数据添加到知识图的新三元组，从而帮助知识图的自动扩展。这两个问题域可以被视为将文本文档中的非结构化数据转换为知识图中的结构化数据的补充技术。拥有结构化格式的数据不仅使人工验证变得更容易，就像自动事实检查一样，而且还允许深入了解其他机器学习模型(如文档分类)如何利用非结构化数据，为可解释的人工智能和模型治理提供了一种途径。

7 Alignment Learning Paradigms
8 对齐学习范式

In this section, we explore the major techniques used in each of the six categories defined above. Each section reviews the works from the perspective of classifying them into supervised, semi-supervised and unsupervised frameworks, motivated by our desire to assess the requisite amount of parallel data necessary to learn an alignment.

在这一节中，我们将探讨在上面定义的六个类别中使用的主要技术。每个部分从将它们分类为监督、半监督和非监督框架的角度来回顾这些工作，这是由我们希望评估学习比对所必需的必要的并行数据量所驱使的。

8.1 Word-to-Word Alignment Techniques
8.2 词对词对齐技术

We proceed by classifying word-to-word alignment techniques into supervised, semi-supervised and unsupervised methods.

我们通过将词到词对齐技术分为监督、半监督和非监督方法来进行。

8.2.1 Supervised Methods
8.2.2 监督方法

Supervised learning methods are the most common and most data intensive in machine learning applications. To help alleviate the burden on developers of these methods, leveraging unsupervised methods as discussed above helps to ingest large amounts of data and build robust features to jumpstart learning. In this section we discuss supervised models that use unsupervised features as inputs with the goal of aligning these resources.

监督学习方法是机器学习应用中最常见和数据最密集的方法。为了帮助减轻这些方法的开发者的负担，利用如上所述的无监督方法有助于摄取大量数据并构建健壮的特征来启动学习。在本节中，我们将讨论使用非监督要素作为输入的监督模型，目标是对齐这些资源。

Regression Models Regression models form the class of solutions first used to address the word-to-word embedding alignment problem. Let us begin by defining languages L_s and L_t, our source and target language, respectively, and embedding functions $f_1 : L_s \rightarrow \mathbb{R}^n$ and $f_2 : L_2 \rightarrow \mathbb{R}^m$. Given a set of parallel translation tokens (w^s, w^t) where
回归模型回归模型形成了首先用于解决单词到单词嵌入对齐问题的一类解决方案。让我们首先定义语言 Ls 和 Lt，分别是我们的源语言和目标语言，并嵌入函数 f1 : Ls → Rn 和 f2 : L2 → Rm。给定一组平行翻译标记(ws，wt)，其中
$w^s_i \in L_s$ and $w^t_i \in L_t$, we wish to learn a transformation matrix W to minimize the following mean-squared loss
$ws_i \in Ls$ 和 $wt_i \in Lt$，我们希望学习一个变换矩阵 W 来最小化下面的均方损失

$$\sum^{n} \left\| W f_1(w^s_i) - f_2(w^t_i) \right\|^2$$

$$\sum_{i=1}^{n} \left\| W f_1(ws_i) - f_2(wt_i) \right\|^2$$

This method was first proposed by [3] as a means of capturing geometric patterns between embeddings across embedding spaces. In the original paper, no additional pre-processing is done on the input word vectors, which were generated using the CBOW word2vec algorithm. The transformation matrix W can then be applied to a new vector $f_1(w_N)$ to map it into the target space where a cosine similarity search can rank all translation candidates. Subsequent papers suggested minor tweaks to the regression model having significant impacts on the capacity to learn. These include the addition of l_2 regularization [47] and adding pre-processing steps such as embedding vector unit normalization, further discussed in the following section.

这种方法最早是由[3]作为捕捉跨嵌入空间的嵌入之间的几何模式的手段。在原始论文中，没有对使用 CBOW word2vec 算法生成的输入单词向量进行额外的预处理。然后可以将变换矩阵 W 应用于新的向量 f1(wN)，以将其映射到目标空间中，在该目标空间中，余弦相似性搜索可以对所有翻译候选进行排序。随后的论文建议对回归模型进行细微调整，这对学习能力有重大影响。这些包括 12 正则化的增加[47]并添加预处理步骤，如嵌入向量单元归一化，这将在下一节中进一步讨论。

Orthogonal Models The original regression model utilized a Euclidean distance in learning the transformation matrix, yet relies on cosine similarity to carry out similarity searches in the target space. This inconsistency was first noted by [48] who in turn modified the regression process to add unit length normalization to the source and target
正交模型原始回归模型在学习变换矩阵时利用欧几里德距离，然而依赖余弦相似性在目标空间中执行相似性搜索。这种不一致首先由[48]他们又修改了回归过程，将单位长度标准化添加到源和目标中

vector spaces and constrain the matrix W to be orthogonal, that is $W^TW = I$ where I is the identity matrix. The pre-processing step and orthogonal constraint then line up with the retrieval method, where we are less concerned with
向量空间，并将矩阵 W 约束为正交的，即 W TW = I 其中 I 是单位矩阵。然后，预处理步骤和正交约束与检索方法一致，这是我们不太关心的
distances between vectors and more concerned with the angles between them. Solutions to this minimization problem are still carried out by stochastic gradient descent where the orthogonality constraint is implemented by solving the SVD problem, typically done by mini-batch fed to the optimizer.
向量之间的距离，更关心它们之间的角度。这个最小化问题的解决方案仍然是通过随机梯度下降来实现的，其中正交性约束是通过解决 SVD 问题来实现的，这通常是通过向优化器馈送小批量来完成的。

Applications of pre-processing and orthogonal constraints spurred further research into ways to manipulate the source and target embedding spaces to further express their geometric structures. In [49] and [50], the authors evaluate several pre- and post-processing steps, building towards a framework of applicable methods. The steps in this framework are enumerated as follows:

预处理和正交约束的应用刺激了对操纵源和目标嵌入空间以进一步表达它们的几何结构的方法的进一步研究。在[49]和[50]，作者评估了几个预处理和后处理步骤，建立了一个适用方法的框架。该框架中的步骤列举如下:

- Normalize the source and target spaces, either using unit norms or mean centering (where each component/feature is forced to have zero mean) as an initial pre-processing step;
- 作为初始预处理步骤，使用单位范数或均值居中(其中每个组件/特征被强制为零均值)来归一化源空间和目标空间；
- Feature whitening, requiring each feature to have unit variance and removing their correlations, applied to both source and target space independently;
- 特征白化，要求每个特征具有单位方差并去除它们的相关性，独立地应用于源空间和目标空间；
- Learning an orthogonal mapping via the regression technique;
- 通过回归技术学习正交映射；
- Re-weight the features to increase their correlations between source and target spaces, only applied if whitening was applied prior to learning the mapping;
- 重新加权特征以增加它们在源空间和目标空间之间的相关性，仅当在学习映射之前应用白化时才应用；
- De-whitening to capture the variance of the original embedding spaces, applied only if whitening was applied prior to learning the mapping; and
- 去白化以捕获原始嵌入空间的方差，仅当在学习映射之前应用白化时才应用；和
- Reducing the dimension by only keeping the most important components of the source and target spaces, helping to remove noise captured in the tail components.
- 通过仅保留源空间和目标空间中最重要的分量来降低维度，有助于移除尾部分量中捕获的噪声。

The authors show that combining these steps helped them achieve superior performance when using CBOW word vectors and 5,000 supervised training examples. The full framework has been packaged and released as open source code under the moniker VecMap, and is used as a baseline in many comparative surveys.

作者表明，结合这些步骤有助于他们在使用 CBOW 单词向量和 5000 个监督训练示例时获得优异的性能。完整的框架已经打包并作为开源代码以 VecMap 的名字发布，并在许多比较调查中用作基线。

Margin Models The methods of the prior two sections rely on variations of mean-squared error to compute and learn from the differences between the source and target space. An alternative modeling technique leverages a max-margin based loss function. These objectives seek to reward the weights associated with positive pairs (in this case, words that are direct translations) while reducing the signal from noise pairs generated either randomly or using a heuristic. In the case of word-to-word translation, the association between pairs is defined by their cosine similarity, thus we may define the max-margin loss as
余量模型前面两部分的方法依赖于均方误差的变化来计算和学习源空间和目标空间之间的差异。另一种建模技术利用基于最大利润的损失函数。这些目标寻求奖励与正对(在这种情况下，是直接翻译的单词)相关联的权重，同时减少来自随机生成或使用启发式方法生成的噪声对的信号。在词到词翻译的情况下，词对之间的关联由它们的余弦相似度来定义，因此我们可以将最大边际损失定义为

$$\sum_{i=1}^{n}\sum_{j=j}^{k} \max\{0, \gamma - \cos(Wf_1(w_i^s), w_i^t) + \cos(Wf_1(w_i^s), w_j^t)\}$$

σ σ max { 0， γ cos(Wf1(ws)， wt) + cos(Wf1(ws)， wt)} i j

where k represents the number of noise pairs (negative samples) and γ is a parameter fixed for setting the margin between positive and negative cases. Using this objective was first proposed by [51] to address issues of hubness seen in regression and orthogonal techniques. The presence of hubs is driven by embeddings that dominate the space due to their high cosine similarity with all other vectors in the source or target space. These hubs can be caused by the overall frequencies of words in the underlying corpus [52], a common mean vector present in all word vectors causing issues of anisotropy in the embedding spaces [53], or issues derived from least-squares regression where low variance points are all grouped together in the target space.

其中 k 表示噪声对(负样本)的数量， γ 是固定的参数，用于设置正样本和负样本之间的界限。使用这个目标是由[51]来解决回归和正交技术中的自大问题。中枢的存在是由支配空间的嵌入驱动的，这是因为它们与源或目标空间中的所有其他向量具有高余弦相似性。这些中枢可能是由底层语料库中单词的总体频率引起的[52]，所有字向量中存在的公共均值向量导致嵌入空间中的各向异性问题[53]，或者从最小二乘回归中导出的问题，其中低方差点都被分组在目标空间中。

Margin-based models are also explored in [54], where the authors also aim to combat the issue of hubs by introducing a new retrieval criteria. The cross-domain similarity local scaling (CSLS) is defined as

基于利润的模型也在[54]，其中作者还旨在通过引入新的检索标准来解决中枢问题。跨域相似性局部缩放(CSLS)被定义为

$$CSLS(x, y) = -2\cos(x, y) + \frac{1}{k}\sum_{y' \in N_Y(x)} \cos(x, y') + \frac{1}{k}\sum_{x' \in N_X(y)} \cos(x', y)$$

CSLS(x， y)= 2 cos(x， y) + 1 k cos(x， y′)+1 k

$$\sum_{x' \in N_X(y)} \cos(x', y) \qquad \sum_{x' \in N_X(y)} \cos(x', y)$$

where $N_Y(x)$ is the set of k nearest neighbors of x in the target space. The authors build this retrieval criteria into their margin model by using unpaired words (those with no explicit translation in the training set) as negative samples when computing nearest neighbors. The full objective function, called the relaxed CSLS (RCSLS) is then computed as

其中 NY (x)是目标空间中 x 的 k 个最近邻居的集合。当计算最近邻时，作者通过使用不成对的词(在训练集中没有显式翻译的词)作为负样本，将这种检索标准构建到他们的边缘模型中。称为松弛 CSLS (RCSLS)的完整目标函数计算如下

$$\frac{1}{n}\sum_{i=1}^{n} -2\cos(Wf(w_i^s), w_i^t) + \frac{1}{n}\sum_{i=1}^{n}\cos(Wf(w_i^s), w_i^t) + \frac{1}{n}\sum_{i=1}^{n}\cos(Wf(w_i^s), w_i^t) + 1$$

$$\sum \cos(Wf(w^s), w) \qquad \sigma \cos(Wf(ws), w)$$

$$\frac{1}{} \sum_{w_j \in N_Y} \qquad \frac{1}{} \sum_{Wf_1(ws_j) \in N_X}$$

For margin-based methods, RCSLS is the most popular method, used as a benchmark for comparisons to other methods [55, 56]. While competitive, we find margin-based methods are studied less frequently; we conjecture this is due to the difficulty in selecting informative negative samples and the preference for methods using as little supervision as possible.

对于基于差值的方法，RCSLS 是最受欢迎的方法，用作与其他方法进行比较的基准[55, 56]. 虽然有竞争力，我们发现基于利润的方法研究较少；我们推测这是由于难以选择信息丰富的阴性样本，以及偏好使用尽可能少监督的方法。

Other Approaches In the interest of exploring methods that extend past word-to-word alignment and are able to generalize to other embedding spaces, we briefly mention alignment methods that lie outside the three categories noted above. The first such method relies on the word neighborhood structures, based on the assumption that for a neighborhood of points in the source space the neighborhood can be reconstructed after applying a linear mapping to the target space. By using manifold learning, without making the assumption that the manifolds (or embedding spaces) were learned via the same algorithm, to capture neighborhood structures, [57] propose a new locality preserving loss function. Given embedding spaces as manifolds M^s and M^t, the goal is to learn a mapping $f : M^s \rightarrow M^t$, which is optimized based on three pieces: an orthogonal piece; a mean-squared error piece; and an additional locality preserving loss piece. This approach represents the base of several models, encapsulating the regression class of models and the orthogonal constrained class of models while adding structure preserving pieces for regularization. The entire loss function can be written as

其他方法为了探索扩展到单词到单词对齐之外并且能够推广到其他嵌入空间的方法，我们简要地提及位于上述三个类别之外的对齐方法。第一种这样的方法依赖于单词邻域结构，基于这样的假设:对于源空间中的点的邻域，在将线性映射应用到目标空间之后，可以重构邻域。通过使用流形学习，而不假设流形(或嵌入空间)是通过相同的算法学习的，来捕捉邻域结构，[57]提出一种新的保局损失函数。给定嵌入空间为流形 Ms 和 Mt，目标是学习映射 f : Ms Mt，其基于三个部分被优化:正交部分；均方误差块；和附加的局部保持损失块。这种方法代表了几个模型的基础，封装了模型的回归类和模型的正交约束类，同时添加了用于正则化的结构保持块。整个损失函数可以写成

$$L = L_{mse}(\theta_f) + \beta L_{lpl}(\theta_f, W) + L_{orth}(W)$$

$$L = Lmse(\theta f) + \beta Llpl(\theta f, W) + Lorth(W)$$

where L_{mse} and L_{orth} are as defined in prior sections and

其中 Lmse 和 Lorth 如前面章节中所定义，并且

$$L_{lpl}$$

Llpl

with β a constant to control the influence of the LPL loss and $m_q^s \in M^s$, $M_q^t \in M^t$. The locality preserving loss was shows to assist in both sentence space and word space alignment, particularly when training dataset sizes are limited.

用 β a常数来控制LPL损失和ms ∈ Ms，Mt ∈ Mt 的影响显示以帮助句子空间和单词空间对齐，特别是当训练数据集大小有限时。

8.2.3 Semi-Supervised Methods
8.2.4 半监督方法

In instances where full parallel corpora or dictionaries are not readily available it is possible to use smaller seed lexicons to build toward larger datasets in a semi-supervised way. One such way of building pseudo-dictionaries' is to identify words that are expressed as the same string in both the source and target language [58]. These typically occur for proper nouns and abbreviations such as FBI and Microsoft. According to [58], this procedure was able to generate nearly 47,000 translation pairs between English and Italian, much larger than the 5,000 most popular terms used by many supervised methods, with excellent evaluated levels of precision.

在不容易获得完全平行的语料库或词典的情况下，可以使用较小的种子词典以半监督的方式构建较大的数据集。构建伪词典的一种方法是识别在源语言和目标语言中被表达为相同字符串的单词[58].这些通常出现在专有名词和缩写词中，如FBI和Microsoft。据[58]，这个过程能够生成近47，000个英语和意大利语之间的翻译对，比许多监督方法使用的5，000个最流行的术语多得多，并且具有出色的评估精确度。

Aside from string matching, other alternative corpus construction methods include using very small seed lexicons (on the order of 25 word pairs) and iteratively adding candidate pairs. The work of [40] propose alternating between a step for learning the mapping W similar to those in Section 4.1.1, followed by a dictionary induction step. Given embedding spaces X and Z, let D be the binary matrix representing the word-pair dictionary between the two languages, i.e. $D_{ij} = 1$ when word i of the source is aligned to word j in the target language. The mapping matrix can then be defined as

除了字符串匹配，其他可选的语料库构建方法包括使用非常小的种子词典(大约25个单词对)和迭代地添加候选词对。的工作40]建议在学习映射W的步骤(类似于第节中的步骤)之间交替4.1.1，接着是字典归纳步骤。给定嵌入空间X和Z，让D是表示两种语言之间的词对词典的二进制矩阵，即当源的词I与目标语言中的词j对齐时，Dij = 1。映射矩阵可以定义为

$$W^* = \arg\min_W \sum_i \sum_j D_{ij} \|WX - Z\|^2$$

At each step of processing, updates to D are computed as $D_{ij} = 1$ if

在处理的每一步，如果

$$j = \arg\max_k (X_{i*} W^{*}) \cdot Z_{k*}$$

$$j = arg\,max_k (Xi \div W \div Zk \div$$

otherwise $D_{ij} = 0$. To evaluate the efficacy of this method, small seed dictionaries are sampled ranging in size from 25 to 2,500 entries, as well as experimenting with only aligning numerals (i.e. digits 0 to 9). Given the limited training set, this self-learning paradigm is competitive with, and at times outperforms, supervised methods.
否则 Dij = 0。为了评估这种方法的有效性，对小种子字典进行采样，其大小范围从 25 到 2，500 个条目，并且只对对齐的数字(即数字 0 到 9)进行实验。给定有限的训练集，这种自我学习范式与监督方法相比具有竞争力，有时甚至优于监督方法。

The work of [59] further explores iterative learning, alternating between supervised alignment and unsupervised distribution matching, as explored in the next section, as well as introducing novel metrics to assess the orthogonality assumptions used in supervised approaches. We further unpack these notions in Section 4.1.3.

的工作 59]进一步探讨了迭代学习，在监督对齐和非监督分布匹配之间交替，如在下一节中所探讨的，以及引入新的度量来评估监督方法中使用的正交性假设。我们在第二节进一步阐述这些概念 4.1.3.

8.2.5 Unsupervised Methods
8.2.6 无监督方法

Under the goal of restricting the amount of parallel data needed to create an alignment between two word spaces, several approaches have been proposed that attempt to leverage the structure of the embedding space itself, completely removing the need for parallel data. A key approach was described in [42] where the authors propose leveraging an adversarial learning paradigm. In this setup, the goal is still to learn a linear map W between the source embedding vectors $f_1(w^s)$ and target space embeddings $f_2(w^t)$. A discriminator D is trained to recognize and separate the mapped

在限制创建两个单词空间之间的对齐所需的并行数据量的目标下，已经提出了几种方法，试图利用嵌入空间本身的结构，完全消除对并行数据的需要。一种关键方法在[42]作者建议利用对抗性学习范式。在这种设置中，目标仍然是学习源嵌入向量 f1(ws)和目标空间嵌入 f2(wt)之间的线性映射 W。鉴别器 D 被训练来识别和分离映射的

embeddings $Wf_1(w_i^s)$ from $f_2(w_i^t)$, while an adversarial generator G is trained to fool D. The given loss functions for
嵌入来自 f2(wt)的 Wf1(ws)，而敌对生成器 G 被训练来愚弄 d

both models are
两种型号都是

$$L(\theta_D | W) = -\frac{1}{n}\sum_{i=1}^{n} \log P(source = 1 | Wx) - \frac{1}{m}\sum_{i=1}^{m} \log P$$

$$L(\theta_D | W) = -\frac{1}{n}\sum_{i=1}^{n} \log P$$

(source = 1 | Wx)1 σ log P

$$P_{\theta_D}(source = 0|y_i)$$

$$P_{\theta_D}(source = 0|y_i)$$

$$\frac{1}{n}\sum_{i=1}^{n}$$

$$\frac{1}{m}\sum_{i=1}^{m}$$

for the discriminator model, and

对于鉴别器模型，以及

$$L(W|\theta_D) = \frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(source = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(source = 0|y_i)$$

$$L(W|\theta_D) = -\frac{1}{n}\sum \log P$$

$$\log P(\text{source} = 0 \mid Wx) \quad (\text{source} = 1 \mid y) \quad \theta_D \quad (\text{source} = 1 \mid y) \quad \theta_D$$

$$n \qquad i \qquad m \; i=1 \qquad i$$

for the generator model. With an initial linear map W learned, the authors then apply a refinement procedure by identifying anchor points as pairs that were frequently identified as translations in the prior step. The anchor points and

对于发电机模型。有了一个初始的线性映射，作者就可以通过将锚点识别为在前面步骤中经常被识别为翻译的对来应用一个改进过程。锚点和

their corresponding word frequencies are used to solve the orthogonal Procrustes problem to generate a refined mapping matrix W^*. This final matrix is used in conjunction with the CSLS objective described in prior sections to mitigate hubness and areas of density in generating a final translation from source to target. The adversarial method has been
它们对应的词频被用于求解正交 Procrustes 问题，以生成精确的映射矩阵 W*。该最终矩阵与前面章节中描述的 CSLS 目标结合使用，以在生成从源到目标的最终转换时减轻中心和密度区域。对抗的方法已经

utilized to generate large benchmark datasets under the name Multilingual Unsupervised or Supervised Embeddings (MUSE), releasing parallel embedding spaces trained using FastText in 110 languages.
用于生成名为多语言非监督或监督嵌入(MUSE)的大型基准数据集，释放使用 110 种语言的 FastText 训练的并行嵌入空间。

As previously discussed, word embedding models tend to reflect the frequency of word usage in the underlying language. While the adversarial method directly leverages word frequencies, an alternative unsupervised method in [41] captures these patterns by analyzing the similarity distributions of the word vectors themselves. By constructing a pair-wise similarity matrix of all word embeddings in the source and target languages, trends in their usages can be exploited to create an initial seed dictionary. By pre-processing to unit normalize the embeddings in both the source X and target Z

如前所述，单词嵌入模型倾向于反映底层语言中单词使用的频率。尽管对抗性方法直接利用词频，但[41]通过分析单词向量本身的相似性分布来捕捉这些模式。通过构建源语言和目标语言中所有单词嵌入的成对相似性矩阵，可以利用它们的使用趋势来创建初始种子字典。通过预处理对源 X 和目标 Z 中的嵌入进行单位归一化
spaces, these similarity matrices can quickly be computed as $M_X = XX^T$ and $M_Y = YY^T$. To further reduce the complexity of finding maps between these similarity matrices, each similarity matrix can then be sorted row by row
空间，这些相似性矩阵可以快速计算为 MX = XXT 和 MY = Y Y T。为了进一步降低在这些相似性矩阵之间查找映射的复杂性，每个相似性矩阵可以逐行排序

to identify the most influential embedding dimension and nearest neighbor searches can then be executed to generate candidate pairs. The seed dictionary can then be expanded using semi-supervised methods described in [40].
为了识别最有影响的嵌入维度，然后可以执行最近邻搜索来生成候选对。然后可以使用中描述的半监督方法扩展种子字典[40].

Aside from leveraging similarity distributions of the underlying embedding spaces, methods are also available to treat these embedding spaces as metric spaces, adopting mathematical tools from measure theory and topology to describe their nature. One such metric is the Gromov–Wasserstein distance used to compare two pairs of spaces, rather than the pairwise point-by-point metrics such as similarities. Using this metric, [43] transform the alignment problem to one of finding an optimal transport from source X to target Z. Due to computational costs, the problem is split into two steps where the two spaces are first aligned using the explicit optimization to find an optimal coupling followed by a refinement using an orthogonal Procrustes procedure, as in [42].

除了利用潜在嵌入空间的相似性分布之外，还可以使用方法将这些嵌入空间视为度量空间，采用来自测度论和拓扑学的数学工具来描述它们的性质。一种这样的度量是用于比较两对空间的格罗莫夫-瓦瑟斯坦距离，而不是像相似性这样的成对逐点度量。使用此指标，[43]将对准问题转化为寻找从源 X 到目标 z 的最佳传输的问题。由于计算成本，该问题被分成两个步骤，首先使用显式优化来对准两个空间，以找到最佳耦合，然后使用正交 Procrustes 过程进行细化，如[42].

8.3 Graph-to-Graph Alignment Techniques

8.4 图形到图形对齐技术

As in word-to-word alignments, graph-to-graph alignment techniques can be classified into supervised, semi-supervised and unsupervised methods. Within each paradigm, however, it is slightly more complicated to develop a straight-forward classification of techniques. We posit this is due not only to the variety of graph datasets available but the velocity at which new research is being published, as noted in Section 3.1. We proceed by categorizing techniques by their level of parallel data needed to learn a robust model. Where applicable, we will also classify techniques according to their approach to the source and target graph embeddings, noting if they utilize translation-based, semantic-matching or graph-structure models.

如同在词到词的对齐中一样，图到图的对齐技术可以分为监督、半监督和非监督方法。然而，在每个范例中，开发一个简单的技术分类稍微复杂一些。我们认为这不仅是由于可用的图表数据集的多样性，也是由于新研究发表的速度，如第节所述 3.1. 我们继续按照学习健壮模型所需的并行数据的级别对技术进行分类。在适用的情况下，我们还将根据它们对源和目标图嵌入的方法对技术进行分类，注意它们是否利用了基于翻译的、语义匹配的或图结构的模型。

8.4.1 Supervised Methods
8.4.2 监督方法

To address issues of coverage in cross-lingual knowledge graphs, [60] propose a method for embedding knowledge graphs in a source and target language and automatically learning alignments between them, called MTransE. Leveraging the translational-based TransE algorithm for generating embeddings of each monolingual knowledge graph G_i and G_j, the embedded triples of each graph are then fed through an alignment scoring function S_a, where the total alignment score is calculated as

为了解决跨语言知识图表的覆盖问题，[60]提出了一种在源语言和目标语言中嵌入知识图并自动学习它们之间对齐的方法，称为 MTransE。利用基于翻译的 TransE 算法来生成每个单语知识图 Gi 和 Gj 的嵌入，然后每个图的嵌入三元组被馈送通过比对评分函数 Sa，其中总比对分数被计算为

$$S_A = \sum_{(T,T') \in \delta(G_i, G_J)} S_a(T, T')$$

$$SA = \sum_{(T, T') \in \delta (Gi, GJ)}$$

$Sa(T,\ T')$

where $\delta_{(G_i,G_j)}$ represents the supervised set of pre-aligned triples. The authors propose three main classes of functions for S_a: distance-based measures, translation vectors and linear transformations. For distance-based measures, the triples in G_i and G_j can be represented as a function of the difference in the head and tail entities

其中 δ (Gi，Gj)表示预对齐三元组的监督集。作者为 Sa 提出了三类主要的函数:基于距离的度量、平移向量和线性变换。对于基于距离的测量，Gi 和 Gj 中的三元组可以表示为头部和尾部实体的差异的函数

$$S_{a_1} = \|h - h'\| + \|t - t'\|$$

$$sa1 \quad \| h \ h' \| + \| t \| t' \|$$

or adjusted to also represent differences in the relation embeddings

或者被调整为也表示关系嵌入中的差异

$$S_{a_2} = \|h - h'\| + \|r - r'\| + \|t - t'\|$$

$$sa2 \quad \| = h \ h \| ' + \| r \ r' \| + t \| t' \|$$

An alternative approach focuses not only on the differences in individual components of the triples, but allows for translation vectors to be learned between the entities and relations, as defined by

另一种方法不仅关注三元组的单个组件的差异，还允许学习实体和关系之间的翻译向量，如所定义的

$$S_{a_3} = \ddot{h} + v^e = h + ve \qquad \underset{v^r}{—} h'\ddot{} + \ddot{r} + \quad —h' + r + VR$$

$Sa3$

— r′′ + t′′ + ve

—r′ — t′′ —t’
+t+ve ij ij ij

where +ve and +vr are learnable translations such that e + ve ≈ e′. The final class of functions defines learnable
其中+ve 和+vr 是可学习的翻译，使得 e+ve≈e’。最后一类函数定义了可学习的
linear transform matrices, with one focused on translations between entities
线性转换矩阵，其中一个专注于实体之间的转换

$$s_{a_4} = M^e_{ij}(h - h')'' + M^e_{ij}(t - t')''$$

$$sa = Me\ h\ h' + Me\ t\ t'$$
4 ij ij

and another with learnable transformations for both entities and
and another with learnable transformations for both

$$s_{a_5} = M^e_{ij}(h - h')'' + M^r_{ij}(r - r')'' + M^e_{ij}(t - t')''$$

$$sa = Me\ h\ h' + Mr\ r\ r' + Me\ t\ t'$$
5 ij ij ij

The authors conclude that the linear transformation models work best with limited differences between the model with entity transformations and the model with both entity and relational translations.

作者的结论是，线性转换模型在具有实体转换的模型和具有实体和关系转换的模型之间的有限差异下工作得最好。

The methods of MTransE lean heavily on the underlying translation model of TransE, which, while directly addressing similarities in graph structures in each individual space, largely ignores other information sources contained in the knowledge graph such as entity types and attributes. Relying only on the structured embedding approaches also

MTransE 的方法严重依赖 TransE 的底层翻译模型，该模型在直接解决每个单独空间中的图结构的相似性的同时，在很大程度上忽略了知识图中包含的其他信息源，例如实体类型和属性。仅仅依靠结构化嵌入方法也

leads to issues when the distribution of relations in the knowledge graph is skewed, as has been widely observed in many large-scale knowledge graphs [61, 62]. To leverage both structure and attributes in aligning both graphs, a joint attribute-preserving embedding (JAPE) module is presented in [63]. For the structure embedding piece, the authors again leverage a translation-based approach letting $f(p) = \|h + r - t\|^2$ where p is a known triple from the supervised training set. They then slightly modify the training process by using a training set P that has pre-aligned triples in both the source and target space to capture correspondences between entities sharing similar relationships. This accomplishes the alignment of both cross-lingual entities and their relations in a single computational step and can be seen as optimizing the score of

当知识图中的关系分布是倾斜的时，会导致问题，正如在许多大规模知识图中广泛观察到的那样[61,62].为了利用结构和属性来对齐两个图，在[63].对于结构嵌入部分，作者再次利用基于翻译的方法，让 f (p) = ‖ h + r t ‖ 2，其中 p 是来自监督训练集的已知三元组。然后，他们通过使用在源空间和目标空间中都具有预先对齐的三元组的训练集 P 来捕捉共享相似关系的实体之间的对应关系，从而稍微修改训练过程。这在单个计算步骤中完成了跨语言实体及其关系的对齐，并且可以被视为优化了

$$L_{SE} = \sum_{p \in P} \sum_{p' \in P'} (f(p) - \alpha f(p'))$$

$$\text{LSE} = \sigma \quad \sigma \quad (\text{f}(\text{p}) \ \alpha \ \text{f}(\text{p}' \ ))$$

$$p \in P \ p' \in P'$$

where α is a tunable margin hyper-parameter and P' a set of negative samples. In addition, attributes of the entities such as their data type and correlations between relation occurrences are used to generate attribute embedding vectors
其中，α 是可调容限超参数，P’是一组负样本。此外，实体的属性，例如它们的数据类型和关系出现之间的相关性，用于生成属性嵌入向量
using a skip-gram like objective function
使用类似跳格的目标函数

$$L_{AE} = - \sum_{(a,c) \in H} w_{a,c} \log p(c|a)$$

$$\text{LAE} = -$$

$$(\text{a}, \ \text{c}) \in \text{H}$$

测井曲线 $p(c|a)$

where H is the set of pairwise positively correlated attributes, i.e. when entity e_j has attribute a it is also highly likely to have attribute c. The attribute embeddings are then used to compute three similarity matrices $S^{(1)}$, $S^{(2)}$, $S^{(1,2)}$ representing the inner-graph entity attribute similarity scores as well as the cross-graph attribute similarity scores. This additional data from the training set helps to build more support for entities to be aligned that can not be captured when using only translational-based models and can be combined to minimize the objective

其中 H 是成对正相关属性的集合，即当实体 ej 具有属性 a 时，它也很可能具有属性 c。然后，属性嵌入被用于计算三个相似性矩阵 S(1)，S(2)，S(1，2)，它们表示图内实体属性相似性得分以及跨图属性相似性得分。来自训练集的这些附加数据有助于为要对齐的实体建立更多支持，这些实体在仅使用基于平移的模型时不能被捕捉，并且可以被组合以最小化目标

$$L_S = \left\| E^{(1)}_{SE} - S^{(1,2)} E^{(2)}_{SE} \right\| + \beta \left(\left\| E^{(1)}_{SE} - S^{(1)} E^{(1)}_{SE} \right\| + \left\| E^{(2)}_{SE} - S^{(2)} E^{(2)}_{SE} \right\| \right)$$

The structured and attribute losses can then be jointly optimized for learning the entire model through

然后，可以联合优化结构化损失和属性损失，以便通过以下方式学习整个模型

$$O_{JOINT} = O_{SE} + \delta O_S$$

OJ OIN T = OSE + δOS

where δ is a tunable hyper-parameter to moderate the influence of the attribute similarities.

其中 δ 是可调的超参数，用于调节属性相似性的影响。

Having demonstrated the importance of incorporating both structure and attributes into the alignment process, other authors followed in the footsteps of the JAPE model, although with different underlying embedding techniques. In the work of [64], the authors utilize the graph convolutional network architecture (GCN-EA) to embed entities from the training sets into an aligned space. By creating a bipartite graph between the pre-aligned entities in the training set the GCN-EA approach models the edges between each distinct graph as equivalence relations, discovering other equivalence relations as alignments by encoding neighborhood information. This approach is then further refined by incorporating the entity attributes in an additional convolutional layer after entity embeddings have been defined. Given two graphs G_1 and G_2 and a training dataset of pairs of matched entities from each, i.e. $S = (e_{m_1}, e_{m_2})$, $e_{m_1} \in G_1$, $e_{m_2} \in G_2$,

在展示了将结构和属性合并到对齐过程中的重要性之后，其他作者也跟随 JAPE 模型的脚步，尽管使用了不同的底层嵌入技术。在...的工作中 64]，作者利用图卷积网络架构(GCN-EA)将来自训练集的实体嵌入到对齐的空间中。通过在训练集中预先对齐的实体之间创建二分图，GCN-EA 方法将每个不同图之间的边建模为等价关系，通过编码邻域信息发现其他等价关系作为对齐。在定义了实体嵌入之后，通过将实体属性合并到附加的卷积层中，该方法被进一步改进。给定两个图 G1 和 G2 以及来自每个图的匹配实体对的训练数据集，即 S = (em1，em2)，em1 G1，em2 G2，

we define two parallel GCN models GCN_1 and GCN_2 to generate embeddings of each input graph. Each GCN

我们定义了两个并行的 GCN 模型 GCN1 和 GCN2 来生成每个输入图的嵌入。每个 GCN

outputs a vector representation of a given input entity, call them v_{m_1} and v_{m_2} for e_{m_1} and e_{m_2}, that can be seen as the concatenation of two parts. The first piece of the output vector v_{m_j} represents the structural piece from the convolutional network with dimension d_s. The second piece represents the attribute representation embedding from the next layer of the network, with dimension d_a. Thus each vector v_{m_j} is of $d_s + d_a$ dimension and compactly represent the structure and attributes of the particular input entity. These representations are then fed to a distance matching function

输出给定输入实体的向量表示，对于 em1 和 em2，将它们称为 vm1 和 vm2，这可以被视为两部分的串联。输出向量 vmj 的第一部分代表来自维度为 ds 的卷积网络的结构部分。第二部分表示从网络的下一层嵌入的属性表示，维度为 da。因此，每个向量 vmj 是 ds + da 维的，并且简洁地表示特定输入实体的结构和属性。这些表示然后被馈送到距离匹配函数

$$D(x, y) = \beta f_s(h_s(x), h_s(y)) + (1 - \beta) f_a(h_a(x), h_a(y))$$

where $f(a, b) = \|a - b\|$, h_s and h_a take the structure and attribute piece of the embedding, respectively, and β is a hyperparameter that balances the trade-off between the importance of structure and attributes. The distances between pre-aligned entities in the training set can then be back-propagated through the network using a margin-based criteria, one for the structure embeddings and one for the attribute embeddings

其中 f（a，b）=‖a b‖，hs 和 ha 分别取嵌入的结构和属性块，β 是一个超参数，它平衡了结构和属性重要性之间的权衡。然后，训练集中预先对齐的实体之间的距离可以使用基于边缘的标准通过网络反向传播，一个用于结构嵌入，一个用于属性嵌入

$$L_s = \sum_{(x,y)\in S} \sum_{(x',y')\in S'} [f(h_s(x), h_s(y)) + \gamma_s - f(h_s(x'), h_s(y'))]_+$$

$$ls = \sigma \sigma [f(hs(x), hs(y)) + \gamma s f(hs(x'), hs(y'))]+$$

$$L_a = \sum_{(x,y)\in S} \sum_{(x',y')\in S'} [f(h_a(x), h_a(y)) + \gamma_a - f(h_a(x'), h_a(y'))]_+$$

$$la = \sigma \sigma [f(ha(x), ha(y)) + \gamma a f(ha(x'), ha(y'))]+$$

with γ_s and γ_a as margin hyper-parameters.

用 γs 和 γa 作为裕度超参数。

Thus far, the three KG-to-KG alignment methods explored, namely MTransE, JAPE and GCN-EA, have focused only on the problem of aligning entities between the two input graphs, and while attribute information from the graphs has also been included, little has been done to also factor in the relations and relational-types in each individual graph. Incorporating relation information is an important facet for selecting an approach to aligning an input source to a knowledge graph embedding, especially in the application to relation extraction from text documents. Equally important in capturing relational data is accounting for directionality; methods must be able to distinguish between one-to-one, many-to-one and many-to-many relation types. Rather than solely relying on aligned entities, [65] create a training set of aligned entities $S_e = (e_{m_1}, e_{m_2}), e_{m_1} \in G_1, e_{m_2} \in G_2$, and aligned relations $S_r = (r_{m_1}, r_{m_2}), r_{m_1} \in G_1, r_{m_2} \in G_2$

到目前为止，所探索的三种 KG-to-KG 比对方法，即 MTransE、JAPE 和 GCN-EA，仅集中在两个输入图之间的实体比对问题上，虽然也包括了来自图的属性信息，但是很少考虑每个单独图中的关系和关系类型。结合关系信息是选择将输入源与知识图嵌入对齐的方法的一个重要方面，特别是在从文本文档中提取关系的应用中。捕捉关系数据同样重要的是考虑方向性；方法必须能够区分一对一、多对一和多对多的关系类型。而不是仅仅依靠联合实体，[65 创建对准实体 Se = (em1，em2)，em1 G1，em2 G2，以及对准关系 Sr = (rm1，rm2)，rm1 G1，rm2 G2 的训练集

in a multi-mapping relation aware technique dubbed MMEA. The authors proceed by defining their own knowledge 在被称为 MMEA 的多重映射关系感知技术中。作者从定义他们自己的知识开始

graph embedding process, avoiding the pitfalls of translation-based methods by defining their own embedding process, called DistMA. DistMA works by replacing translation-based distances with inner-products, defined by

图形嵌入过程，通过定义自己的嵌入过程来避免基于翻译的方法的缺陷，称为 DistMA。DistMA 的工作原理是用内积代替基于平移的距离，内积定义为

$$E_1(h, r, t) = (v_h, v_r) + (v_r, v_t) + (v_h, v_t)$$

$$E1(h, r, t) = (vh, vr) + (vr, vt) + (vh, vt)$$

and replace the margin-based optimization with a logistic loss function, defined by

并且用逻辑损失函数代替基于边际的优化，逻辑损失函数定义为

$$-\sum_{(h,r,t)\in S^{+}} \log \sigma(E_1(h,r,t)) - \sum_{(h',r',t')\in S^{-}} \log \sigma(-E_1(h',r',t'))$$

对数 σ（E1（h，r，t））

(h，rσ，t)

∈S+

(h ’，r ’，t’）∈S

$\log \sigma(-1 \cdot E_1(h', r', t')) + \lambda \quad \theta$ E1(h’，r’，t’)+λ θ

对数 σ（1 ‖ ‖

Using the inner product rather than subtraction-based distances allows the embeddings to scale well to multi-relational facts in the graph, where methods like TransE typically struggle. The downside is that the proposed DistMA is highly symmetric, making no distinction between the head and tail of the triple, thus incorporating no sense of the directionality of the relation. To combat this, the authors also leverage the ComplEx embedding method, previously presented in Section 2.3.2. Letting

使用内积而不是基于减法的距离允许嵌入很好地扩展到图中的多关系事实，而像 TransE 这样的方法通常很难做到。缺点是提出的 DistMA 是高度对称的，没有区分三元组的头和尾，因此没有包含关系的方向性。为了解决这个问题，作者们还利用了复杂的嵌入方法，这在前面的第节中介绍过 2.3.2. 出租

$$E_2 = \square(|w_h, w_r, w_t|), w_i \in C$$

$$E2 = \quad (\text{(wh，西印度群岛，wt)}), wi \in C$$

the final scoring function for each triple is combined and written as

每个三元组的最终得分函数被组合并写成

$$E(h, r, t) = E_1(h, r, t) + E_2(h, r, t)$$

$$E(h，r，t) = E1(h，r，t) + E2(h，r，t)$$

With both entities and relations embedded for each knowledge graph, the airs from the training set are aligned in a common embedding space such that their representations are equal. This is accomplished by using a cosine similarity metric

通过为每个知识图嵌入实体和关系，来自训练集的 air 在公共嵌入空间中对齐，使得它们的表示是相等的。这是通过使用余弦相似性度量来实现的

$$sim(e_i, e_j) = \|v_{e_i}\|, \ddot{v}_{e_j}\ddot{}\rangle$$

$$(\text{(t)}ve_i\|, \lambda\|$$

to build a similarity matrix $S_{1,2}$ between the two knowledge graphs. This similarity matrix can then be ranked from both directions, i.e. for the similarities $M_{1,2} : G_1 \rightarrow G_2$ and $M_{2,1} : G_2 \rightarrow G_1$. The final ranking matrix is then computed as $M = M_{1,2} + M_2{}^{\mathsf{T}}{}_{,1}$.

以在两个知识图之间建立相似性矩阵 1，2。然后可以从两个方向对该相似性矩阵进行排序，即对于相似性 M1，2 : G1 → G2 和 M2，1 : G2 → G1。最终的排序矩阵然后被计算为 M = M1，2 + M2T，1。

In continuation of research on addressing multi-relational patterns, [66] focus on Non-Translational Alignment for Multi-relational networks (NTAM). Rather than relying on a semantic energy, translational-based, or graph convolutional model to build embeddings, the authors build a probabilistic model based on *motifs* that can be found within the graph. These motifs, or graph patterns, include triangular structures in the graph where a given node in the triangle can have in-degree of zero (out-degree two), one (out-degree one) or two (out-degree zero), as is accomplished in [67]. While these motifs are flexible in capturing local structures in the graph, aligning these structures required nodes in each individual graph to exhibit very similar neighborhood structures, an assumption that may not hold in large-scale heterogeneous graphs.

在继续关于多关系模式的研究中，[66]关注多关系网络的非翻译比对(NTAM)。作者没有依赖语义能量、基于翻译或图卷积模型来构建嵌入，而是基于图中可以找到的基序来构建概率模型。这些图案或图形模式包括图形中的三角形结构，其中三角形中的给定节点可以具有 0 度的入度(2 度的出度)、1 度的入度(1 度的出度)或 2 度的出度(0 度的出度)，如[67].虽然这些基序可以灵活地捕捉图中的局部结构，但排列这些结构需要每个单独图中的节点表现出非常相似的邻域结构，这种假设在大规模异构图中可能不成立。

For the majority of knowledge graph alignment approaches discussed above, the underlying embedding algorithms rely on negative samples, or false facts, to be generated. These negative sampling paradigms inherently make use of the closed-world assumption wherein all facts are assumed to be contained in the knowledge graph. In opposition is the open-world assumption, where we have only build a knowledge graph of our currently known facts, and the validity of those not contained in this set is uncertain. Using negative sampling instantiates a closed-world assumption, and when those negative facts turn out to actually be true, model performance suffers. In opposition, adversarial networks leverage two networks that attempt to trick one another. The first network, the generator, attempts to create samples that look similar to those in the original data distribution, yet are created in a synthetic way. The job of the second network, the discriminator, is to differentiate between instances of the true dataset versus those coming from the generator. Adversarial networks have been utilized to generate embeddings of single knowledge graphs [68] and can also be used in aligning the representations of distinct knowledge graphs. Based on the notion that the embedding spaces of each graph should have similar spatial features for entities that are likely the same, called by the authors of [69] the embedding distribution, an adversarial network can be used to learn to discriminate between these embedding distributions in order to learn an approximate isomorphism between the two spaces. To accomplish this task, the authors introduce the representation module, the mapping module and the adversarial module. In the representation module, two separate instances of the TransE model are trained on each graph, creating two sets of embeddings e_s and e_t. These embedding matrices are then fed into the mapping module, where seed pairs $S = (e_{s_i}, e_{t_i})$ are used to learn a linear mapping, defining the loss function as

对于上面讨论的大多数知识图对齐方法，底层嵌入算法依赖于要生成的负样本或虚假事实。这些负采样范例固有地利用了封闭世界假设，其中所有事实都被假设为包含在知识图中。与之相对的是开放世界假设，在这种假设中，我们只构建了一个我们当前已知事实的知识图，而那些不包含在这个集合中的事实的有效性是不确定的。使用负采样实例化了一个封闭世界的假设，当那些负事实被证明是真的时，模型性能就会受到影响。相反，敌对网络利用两个试图欺骗对方的网络。第一个网络是生成器，它试图创建看起来类似于原始数据分布中的样本，但却是以合成方式创建的。第二个网络(鉴别器)的工作是区分真实数据集的实例和来自生成器的实例。敌对网络已经被用来生成单一知识图的嵌入[68]并且还可以用于对齐不同知识图的表示。基于每个图的嵌入空间对于可能相同的实体应该具有相似的空间特征的概念，由[69 嵌入分布，可以使用对抗网络来学习区分这些嵌入分布，以便学习两个空间之间的近似同构。为了完成这个任务，作者引入了表示模块、映射模块和对抗模块。在表示模块中，在每个图上训练 TransE 模型的两个独立实例，创建两组嵌入 es 和 et。这些嵌入矩阵然后被馈送到映射模块，其中种子对 S = (esi，eti)用于学习线性映射，将损失函数定义为

$$L_M = \sum_{(e_s,e_t)\in S} \sum_{(es,\,et)\in S} Ge_s - e_t$$

$$LM =$$

ges et ‖ ‖

As in approaches for word-to-word embedding alignment, G can be restricted to be an orthogonal matrix. The authors introduce two additional constraints: the feature reconstruction constraint and the mapping reconstruction constraint. Intuitively, the feature reconstruction constraint dictates that once an embedding for an entity is mapped from the source space to the target space, that same mapping can be applied to map it back to the source space representation. This can be reflected in the adjusted loss function

如同在单词到单词嵌入对齐的方法中一样，G 可以被限制为正交矩阵。作者引入了两个附加约束:特征重构约束和映射重构约束。直观地，特征重构约束规定，一旦实体的嵌入从源空间映射到目标空间，相同的映射可以被应用来将其映射回源空间表示。这可以反映在调整后的损失函数中

$$L'_M = \sum_{(e_s,e_t)\in S} \lambda_1 \left\| Ge_s - e_t \right\| + \lambda_2(\mu \left\| e_s - G^\top Ge_s \right\| + (1-\mu) \left\| e_t - GG^\top e_t \right\|)$$

$$I'_M = \sum_{(es,\,et)\in S}$$

$$\lambda 1\ Ges\ et+\lambda 2(es\ GTGes+(1)et\ gg\ Tet)$$

where λ_1, λ_2 are learnable weights to balance the reconstruction constraint and μ is a harmonic factor. The mapping reconstruction constraint is a variant on restricting G to be orthogonal, forcing the learning algorithm to push G toward the nearest orthogonal manifold, modifying the loss function as

其中 λ1、λ2 是用于平衡重构约束的可学习权重，并且是谐波因子。映射重构约束是将 G 限制为正交的变体，迫使学习算法将 G 推向最近的正交流形，将损失函数修改为

$$L'' = \sum_{(e_s,e_t)\in S} \sum_{(es,\ et)\in S}$$

$$\lambda_1 \left\| G e_s - e_t \right\| + \lambda_2 \left\| G^{\top} G - E \right\|_F \qquad \lambda_1 \left\| G e_{source} - e_{target} \right\| + \lambda_2 \left\| G^{\top} G - E \right\|_F$$

where E is the identity matrix and F is the Frobenius norm. With a mapping learned, the mapped embeddings can be fed to the adversarial module where their synthetic counterparts are build by a generator network while the discriminator network learns to differentiate between true and false examples. The authors show that the adversarial setup helps with generalization as its main focus is on aligning the topological features of each embedding space in a way which reduces sensitivity to noise.

其中 E 是单位矩阵，F 是 Frobenius 范数。随着映射的学习，映射的嵌入可以被馈送到对抗模块，其中它们的合成对应物由生成器网络构建，而鉴别器网络学习区分真和假的例子。作者表明，对立设置有助于推广，因为它的主要焦点是以降低对噪声的敏感性的方式排列每个嵌入空间的拓扑特征。

8.4.3 Semi-Supervised Methods

8.4.4 半监督方法

While the issue of building supervised word-to-word datasets is a challenge for word alignment techniques, the issue is even more prevalent in knowledge graphs due to their large degree of heterogeneity. Building links between entities in disparate knowledge graphs often requires the intervention of human experts and comes at a significant cost. Rather than rely solely on labeled instances, semi-supervised approaches build from a set of seed aligned entities, iteratively building confidence in newly aligned pairs and expanding the set of training examples.

虽然构建有监督的词到词数据集的问题对于词对齐技术来说是一个挑战，但是由于知识图的高度异质性，该问题在知识图中甚至更加普遍。在不同的知识图中的实体之间建立链接通常需要人类专家的干预，并且成本很高。半监督方法从一组种子对齐的实体开始构建，迭代地在新对齐的对中建立置信度，并扩展训练样本集，而不是仅仅依赖于标记的实例。

By embedding the separate graphs using translational-based methods, the authors of [70] build a joint embedding space and utilize a soft alignment scoring function to estimate a reliability score of aligned entities. These three modules are trained in an iterative fashion, making updates to the training set and the joint embeddings at each step. For the individual embeddings, the authors utilize the path-inclusive embeddings of PTransE [37], generating two entity embedding sets E_1 and E_2. To build a joint embedding space, the authors propose three methods: a translation-based model, a linear model and a parameter sharing model. The translation-based model, IPTransE, introduces a new alignment relation r that maps $e_s \in E_1$ to $e_t \in E_2$, where

通过使用基于平移的方法嵌入单独的图,[70 构建联合嵌入空间，并利用软对齐评分函数来估计对齐实体的可靠性评分。这三个模块以迭代的方式被训练，在每一步对训练集和联合嵌入进行更新。对于单个嵌入，作者使用 PTransE [37]，生成两个实体嵌入集 E1 和 E2。为了构建联合嵌入空间，作者提出了三种方法:基于平移的模型、线性模型和参数共享模型。基于翻译的模型 IPTransE 引入了将 es ∈ E1 映射到 et ∈ E2 的新对齐关系 r，其中

$$E(e_s, e_t) = \| e_s + r^{(E_1 \to E_2)} - e_2 \|$$

$$E(es, et) = \|es + r(E1 \to E2) - E2\|$$

The linear mapping model replaces this relation with a transformation matrix M such that

线性映射模型用变换矩阵 M 代替这种关系，使得

$$E(e_s, e_t) = \| M^{(E_1 \to E_2)} e_1 - e_2 \|$$

$$E(es, et) = \|M(E1 \to E2)e1e\ 2\|$$

The parameter sharing attempts to make no mapping, instead forcing aligned entities from the seed set L to have the same embedding representation, such that

参数共享试图不进行映射，而是迫使来自种子集 L 的对齐实体具有相同的嵌入表示，使得

$$e_s \equiv e_t, (e_s, e_t) \in L$$

$$es \equiv et, (es, et) \in L$$

Once entities are projected into a joint space, each entity embedding in the source space is compared to all entity embeddings in the target space, building an aligned entity where

一旦实体被投影到联合空间中，嵌入在源空间中的每个实体与嵌入在目标空间中的所有实体进行比较，构建对齐的实体，其中

$$e\hat{}_t = \arg\min(E(e_s, e_t)), E(e_s, e\hat{}_t) < \theta$$

$$E\check{}t = arg\ min(E(es,\ et)),\ E(es,\ E\check{}t) < \theta$$

where θ is a hyperparameter controlling the distance. The pair $(e_s, e\hat{}_t)$ can then be added to L and the joint alignments can be updated accordingly. In addition to simply updating the set L, the authors also introduce a soft alignment function

其中 θ 是控制距离的超参数。然后可以将对(es，e ˇ t)添加到 L，并且可以相应地更新关节对齐。除了简单地更新集合 L 之外，作者还引入了软对齐函数

$$R(e_s, e_t) = \sigma(k(\theta - E(e_s, e_t))$$

$$R(es,\ et) = \sigma\ (k(\theta E(es,\ et))$$

that tracks the reliability of the new pair, where σ is the softmax function and k is a tunable hyperparameter.

它跟踪新对的可靠性，其中 σ 是 softmax 函数，k 是可调超参数。

Also leveraging translational-based models for embedding each knowledge graph, semi-supervised entity alignment with degree differences (SEA) [71] adjusts the TransE approach by incorporating information about each entities degree when building the embeddings. The key insight is that entities that are well-connected appear in more triples, and thus have more robust embeddings containing more information than entities that are infrequently occurring in the graph. The more frequently occurring entities thus form hubs in each embedding space, making the alignment maps learned biased toward these points. Adjusting TransE to better reflect the degree distribution of each entity helps to alleviate these issues and build more robust alignment maps. To prevent entities with similar degree from clustering together in the embedding space, the authors use an adversarial network where a generator builds degree-aware embeddings while two discriminators D_1 and D_2 are used to classify entities with high or normal degree and entities with low degree, respectively. By designing the generator to create high-quality embeddings to fool the discriminators, those embeddings will encode the entities degree in such a way that embeddings of various degrees become linearly inseparable, and thus don't occupy a dense area of the embedding space. The adversarial training is done by training the TransE representations with the discriminators fixed, then alternating by training the discriminators with the embeddings fixed, generating two sets of degree-aware KG embeddings θ^1 and θ^2. To align the entities from the set of pre-labeled seed

还利用基于翻译的模型来嵌入每个知识图，具有程度差异的半监督实体对齐(SEA) [71]通过在构建嵌入时结合关于每个实体程度的信息来调整转换方法。关键的见解是，良好连接的实体出现在更多的三元组中，因此比图中不常出现的实体具有包含更多信息的更健壮的嵌入。因此，更频繁出现的实体在每个嵌入空间中形成中枢，使得所学习的比对图偏向这些点。调整 TransE 以更好地反映每个实体的程度分布有助于缓解这些问题并建立更健壮的比对图。为了防止具有相似度的实体在嵌入空间中聚集在一起，作者使用了一个敌对网络，其中生成器构建度感知嵌入，而两个鉴别器 D1 和 D2 用于分别对具有高或正常度的实体和具有低度的实体进行分类。通过设计生成器来创建高质量的嵌入以欺骗鉴别器，这些嵌入将以各种程度的嵌入变得线性不可分的方式对实体程度进行编码，从而不会占据嵌入空间的密集区域。对抗训练通过训练具有固定鉴别器的转换表示来完成，然后通过训练具有固定嵌入的鉴别器来交替进行，产生两组度感知 KG 嵌入 θ1 和 θ2。为了从预先标记的种子集合中对齐实体

图/示 jj

alignments L, cycle consistent translation matrices are learned by minimizing

对齐 L，循环一致的翻译矩阵通过最小化来学习

$$\sum \sigma \; M \;\; 1 \;\; \theta \;\; 1 \quad \left\| M^{1}\theta^{1} - \theta_{2} \right\|_{2}^{2} + \left\| M_{2} \; 2 - \theta \right\|_{2}^{2} + M \;\; \theta$$

— θ ¨

— θ

e_i e_j

$(e_i,e_j)\in L$ (ei, ej) ∈L

where the cycles
那里的周期

e_i e_j $\boldsymbol{e_j}$ $\boldsymbol{e_i}$ e_j e_i

$$\theta^1_{\boldsymbol{e_i}} \rightarrow M^1\theta^1_{\boldsymbol{e_i}} \rightarrow M\ M\ \theta_{\boldsymbol{e_i}}$$

$$e_i \overset{\theta^1}{\rightarrow} M\ \theta^1 e_i$$

$$\theta^2_{\boldsymbol{e_j}} \rightarrow M^2\theta^2_{\boldsymbol{e_j}}$$

$$e_j \overset{\theta^2}{\rightarrow} M\ \theta^2 e_j$$

→ M M θei → M M θej

help to improve generalizability to unlabeled instances. By using all the unlabeled entities in generating the degree aware embeddings, the SEA model is able to leverage both labeled and unlabeled entities in building a robust alignment.

有助于提高对未标记实例的推广能力。通过在生成程度感知嵌入中使用所有未标记的实体，SEA 模型能够在构建鲁棒的比对中利用标记的和未标记的实体。

By incorporating a bootstrapping approach, the authors of [72] completely abandon the translational-based model and opt instead for a margin-based model designed to directly leverage information contained in the positive and negative samples sets, which are continuously expanded as the model trains. For triples τ, the objective function

通过引入自举方法,[72]完全放弃基于翻译的模型，转而选择基于边界的模型，该模型旨在直接利用正样本集和负样本集中包含的信息，正样本集和负样本集随着模型训练而不断扩展。对于三元组 τ，目标函数

$$O_e = \sum_{\tau' in T^-} [f(\tau) = \gamma_1] + \mu_1 \sum [\gamma_2 - f(\tau')]$$

OE = σ [f (τ) = γ 1] + 1 σ
τ τ′

[γ2 f(τ’)]

where T^+, T^- refer to the sets of positive and negative triples, respectively. In building T^+ and T^-, the authors introduce an ϵ-truncated negative sampling strategy, emphasizing that corruption of the head or tail entity should be done so in an intelligent way to maximize the signal the model can learn from. In this negative sampling paradigm, the negative candidates are selected from a neighborhood of $s = (1\ \epsilon)N$ based on the cosine similarity of their embeddings.

其中 T+，T-分别指正三元组和负三元组的集合。在构建 T+和-t-时，作者引入了 ϵ-truncated 负采样策略，强调应该以智能的方式来破坏头部或尾部实体，以最大化模型可以学习的信号。在这种负采样范例中，负候选是基于它们嵌入的余弦相似性从 s =(1)ϵ)n 的邻域中选择的。

After t steps of training, the set T^+ is updated based on the bootstrapping procedure. As these new bootstrapped instances may contain errors, the authors introduce an editing technique to dampen their effect. Prior to new candidates y and y' with a truth label x being added to T^+, they are evaluated through

在 T 步训练之后，基于自举过程更新集合 T+。由于这些新的引导实例可能包含错误，作者引入了一种编辑技术来抑制它们的影响。在具有真值标签 x 的新候选 y 和 y’被添加到 T+之前，它们通过

$$\Delta^{(t)}_{(x,y,y')} = \pi(y|x;\Theta^{(t)}) - \pi(y'|x;\Theta^{(t)})$$

$\Delta^{(t)}_{(x,\ y\ ,\ y')}$

$$= \pi(y \mid x; \theta(t)) \rightarrow \pi(y' \mid x; \theta(t))$$

to determine the highest likelihood of the label, preventing uncertainty from leaking into the bootstrapped training set.

以确定标签的最高可能性，防止不确定性泄漏到引导训练集中。

8.4.5 Unsupervised Methods
8.4.6 无监督方法

While there is limited research in semi-supervised methods for knowledge graph alignment, fully unsupervised methods are even less common. In their survey of the literature, [8] claim to observe no research articles on unsupervised methods for knowledge graph alignment. In the time between the release of the survey and this publication, we have found an example of authors exploring unsupervised techniques. In [73] an adversarial training paradigm is used to build links between two graphs. The embeddings of each graph are generated using the DeepWalk technique, taking advantage of the structural properties of each individual graph. These are then mapped using a matrix W fed to a discriminator to differentiate between the source and target space. Given the recency of this publication, its current lack of peer review, and experimentation using only social network domains, we leave the other details to the reader but include its mention to highlight an area of growing interest for researchers.

虽然对知识图对齐的半监督方法的研究有限，但完全无监督的方法更不常见。在他们对文献的调查中 8]声称没有观察到关于知识图对齐的无监督方法的研究文章。在调查发布和本出版物之间的时间里，我们发现了一个作者探索无监督技术的例子。在[73 对抗性训练范例用于在两个图之间建立链接。每个图的嵌入都是使用深走技术生成的，利用了每个单独的图的结构属性。然后使用矩阵 W 将这些映射到鉴别器，以区分源空间和目标空间。鉴于这份出版物的新近性，其目前缺乏同行审查，以及仅使用社交网络领域的实验，我们将其他细节留给读者，但包括其提及，以突出研究人员越来越感兴趣的领域。

8.5 Sentence-to-Sentence Alignment Techniques
8.6 句子到句子的对齐技术

While many of the techniques applied to word-to-word alignment also apply to sentences, there is also a line of research that focuses only on techniques for sentence alignment. Sentence alignment introduces an additional complexity over word alignment due to variability in word ordering and syntactic and morphological differences between languages that challenge the efficacy of traditional mapping based systems. Applications in this space tend to focus on applications to neural machine translation (NMT), however, we believe that these techniques have applications to other research domains.

虽然许多应用于词到词对齐的技术也适用于句子，但是也有一些研究只关注句子对齐的技术。由于单词排序的可变性以及语言之间的句法和形态差异，句子对齐比单词对齐引入了额外的复杂性，这挑战了传统的基于映射的系统的效率。该领域的应用倾向于集中在神经机器翻译(NMT)的应用上，然而，我们相信这些技术也可以应用于其他研究领域。

8.6.1 Supervised Methods

8.6.2 监督方法

While there are a limited number of publications exploring supervised sentence to sentence embedding alignment, there are applications of existing word-to-word techniques to this domain. In [74], to benchmark their semi-supervised method the authors re-implement several alignment models for the purpose of mapping sentence embeddings for machine translation. Specifically, they use the linear regression model model from [3], the l_2 regularized model of [47], and the inverted softmax model of [58]. Details of each of these approaches are given in Section 4.1.1. In their evaluation of these techniques, the authors find that these simple linear techniques and their extensions outperform the more advanced models from seq2seq [75], fairseq [76] and LASER [27]. As these more advanced models do not perform explicit alignments, we leave these for the reader to explore.

虽然有有限数量的出版物探索监督的句子到句子嵌入对齐，但是存在现有的词到词技术在该领域的应用。在[74]，为了测试他们的半监督方法，作者重新实现了几个对齐模型，用于映射机器翻译的句子嵌入。具体来说，他们使用线性回归模型模型从[3]，12 正则化模型的[47]，以及反转的 softmax 模型[58].这些方法的细节在第节中给出 4.1.1. 在对这些技术的评估中，作者发现这些简单的线性技术及其扩展优于 seq2seq [75]，fairseq [76]和激光[27].由于这些更高级的模型不执行显式比对，我们将这些留给读者去探索。

8.6.3 Semi-Supervised Methods
8.6.4 半监督方法

With an eye toward reducing the amount of parallel data necessary, [74] utilize bidirectional GANs for aligning sentence representations. In addition to defining a piece of the loss function for sentence representation pairs (x, y) that are in the training set, the authors also use all available non-parallel data in their approach. The resulting objective function

着眼于减少必要的并行数据量，[74]利用双向 GANs 来对齐句子表示。除了为训练集中的句子表示对(x，y)定义一个损失函数之外，作者还在他们的方法中使用了所有可用的非平行数据。得到的目标函数

$$L_{real} = E_{x,y}[\log(D_{real}(x, y)] + E_x[\log(1 - D_{real}(x, G_X(x)))] + E_y[\log(1 - D_{real}(y, G_Y(y)))]$$

Lreal = Ex，y[log(Dreal(x，y)]+Ex[log(1 Dreal(x，GX(x))))]+Ey[log(1 Dreal(y，GY(y))]

where a real pair (x, y) in the parallel set is contrasted with fake pairs $(x, G_X(x))$ and $(y, G_Y(y))$ created by the respective generators for the source and target space. To further leverage the data available from the non-parallel sentences, the authors additionally introduce two loss functions. The first is designed to minimize the expected value of

其中平行组中的真实对(x，y)与由源和目标空间的相应生成器创建的伪对(x，GX (x))和(y，GY (y))形成对比。为了进一步利用来自非平行句子的可用数据，作者额外引入了两个损失函数。第一种方法旨在最小化的期望值

mismatched pairs, or negative samples $(x', y') \in X \times Y$ such that

不匹配对或负样本(X′，Y′)∈X×Y，使得

$$L_{mis} = E_{x',y'}[\log(1 - D_{real}(x', y'))]$$

Lmis = Ex′，y′[log(1d real(x′，y′)]

The second loss function added includes an additional discriminator to distinguish whether the sentence embedding came from the source or target space, defined as

添加的第二个损失函数包括一个额外的鉴别器，用于区分句子嵌入是来自源空间还是目标空间，定义为

$$L_{dom} = E_x[log(D_{dom}(x, G_X(x)))] + E_y[log(1 - D_{dom}(y, G_Y(y)))]$$
Ldom = Ex[log(Ddom(x, GX(x)))]+Ey[log(1 Ddom(y, GY(y)))]

These three loss functions contribute equally to the overall model loss

这三个损失函数对整体模型损失的贡献相等

$$L = L_{real} + L_{mis} + L_{dom}$$
L = Lreal + Lmis + Ldom

To represent the sentences in both the source and target space, the authors use FastText word vectors and simply average each word representation to create a sentence embedding. In their ablation study, the authors additionally experiment with TF-IDF weighting, finding that for corpora with longer sentence length the TF-IDF weighting leads to improved accuracy, while shorter sentences do not exhibit similar gains. The intuition behind these findings is that much of the noise in longer sentences, such as stop words and other semantically irrelevant words used for syntactic purposes, can be down-weighted, thus creating a more semantically meaningful sentence representation.

为了表示源空间和目标空间中的句子，作者使用 FastText 单词向量，并简单地平均每个单词表示来创建句子嵌入。在他们的消融研究中，作者另外实验了 TF-IDF 加权，发现对于句子长度较长的语料库，TF-IDF 加权导致准确性提高，而较短的句子没有表现出类似的增益。这些发现背后的直觉是，较长句子中的许多干扰，如停用词和其他用于句法目的的语义无关的词，可以被向下加权，从而创建更有语义意义的句子表示。

8.6.5 Unsupervised Methods
8.6.6 无监督方法

Due to the added complexities of mapping full sentence representations, there is limited research on completely unsupervised sentence embedding mappings. For that reason, we restrict our evaluation to the methods demonstrated in [77]. To address the issue of lack of parallel corpora for supervised alignment, the authors of [77] introduce the notion of interlingual semantic representations (ISR) for the few or zero-shot cases. ISR attempts to create an intermediate, low-dimensional space that captures word and sentence semantics that can be fine-tuned to any language or downstream task. To accomplish this representation, the authors utilize an adversarial approach, building a sequence of generators and discriminators to encode language into intermediate representations using only a monolingual corpus. A single generator G makes use of an encoding step *enc* to map an input sentence s to ISR, then applies a decoding step *dec* back to a translated form of that sentence $s\hat{}$. The translation is then fed to the discriminator D for the dual task of determining if the sentence translation is real or synthetic, as well as making a classification for the language of that translation.

由于映射完整句子表示的额外复杂性，对完全无监督的句子嵌入映射的研究有限。因此，我们的评估仅限于[77].为了解决监督比对缺乏平行语料库的问题,[77]引入语际语义表征(ISR)的概念，用于少数或零命中率的情况。ISR 试图创建一个中间的低维空间来捕捉单词和句子的语义，这些语义可以根据任何语言或下游任务进行微调。为了实现这种表示，作者采用了对抗的方法，建立了一系列的生成器和鉴别器，仅使用单语语料库将语言编码为中间表示。单个生成器 G 利用编码步骤 enc 将输入句子 s 映射到 ISR，然后将解码步骤 dec 应用回该句子 s’的翻译形式。然后，该翻译被馈送到鉴别器 D，用于确定句子翻译是真实的还是合成的，以及对该翻译的语言进行分类的双重任务。

Simultaneously, the translation $s\hat{}$ is backpropagated through the generator G for calculation of two losses. The first loss, called the ISR loss, minimizes the distance between the ISR of the forward translation and the backward translation, helping to fine tune the intermediate embedding layer. The second loss, the reconstruction loss, measures how closely the original input sentence vector s and the forward-backward representation after two applications of the generator $G(G(s))$ match. This type of cycle-consistency loss has been shown to aid in cross-domain translation, as in [78], and the authors claim that cycle-consistency helps to preserve semantic information flow from the source to intermediate representations. For fixed embedding inputs, the authors use a BERT-as-a-service model for generation, train their neural architecture using the XNLI dataset, and demonstrate through an ablation study the importance of the reconstruction loss, showing that the cycle-consistency constraint does help in the zero-shot learning task. Fully unsupervised sentence translation remains a challenging yet open research problem, and the pace of research publications in this area continues to increase.

同时，平移 s’通过发电机 G 反向传播，用于计算两个损耗。第一个损失称为 ISR 损失，它最小化正向转换和反向转换的 ISR 之间的距离，有助于微调中间嵌入层。第二个损失，重建损失，测量在两次应用生成器 G(G(s))之后，原始输入句子向量 s 和向前-向后表示有多接近地匹配。这种类型的循环一致性损失已被证明有助于跨域翻译，如[78]，作者声称循环一致性有助于保持从源到中间表示的语义信息流。对于固定的嵌入输入，作者使用 BERT-as-a-service 模型进行生成，使用 XNLI 数据集训练他们的神经架构，并通过消融研究证明重建损失的重要性，表明循环一致性约束确实有助于零炮学习任务。完全无监督的句子翻译仍然是一个具有挑战性但开放的研究问题，在这一领域的研究出版物的步伐继续增加。

9 Benchmark Datasets
10 基准数据集

In this section, we document some of the most popular datasets used in the alignment literature.

在本节中，我们记录了比对文献中使用的一些最流行的数据集。

10.1 Cross-language Word Alignment Benchmarks
10.2 跨语言单词对齐基准

As introduced in Section 1.1, the task of Bilingual Lexical Induction aims to evaluate the consistency and ability to learn alignments between embeddings of two distinct languages. There are two themes that datasets in this space may be classified as: those that provide aligned source and target text (akin to datasets used for machine translation) and those that provide pre-trained embeddings of the source and target languages along with seed alignments. In regards to the latter, such published datasets typically select a base embedding algorithm (i.e. FastText or GloVe) and generate pre-trained embeddings on several monolingual corpora, each using the same model hyper-parameters to dictate consistency.

如第节所述 1.1，双语词汇归纳的任务旨在评估两种不同语言嵌入之间的一致性和学习对齐的能力。这个空间中的数据集可以分为两个主题:提供对齐的源和目标文本的主题(类似于用于机器翻译的数据集)，以及提供源和目标语言的预训练嵌入以及种子对齐的主题。关于后者，这种公布的数据集通常选择基本嵌入算法(即 FastText 或 GloVe)并在几个单语语料库上生成预训练嵌入，每个使用相同的模型超参数来规定一致性。

A modern and oft-cited toolkit for BLI datasets is the Facebook MUSE dataset [42]. This dataset contains monolingual embeddings and seed dictionaries for 30 languages, as well as bilingual seed dictionary pairs for 110 languages. For languages coupled with monolingual embeddings, all such embeddings were generated using the FastText algorithm trained over a copy of Wikipedia in the respective language. Each set of embeddings has been generated using the Skip-gram model with the embedding dimension set to 300 with no additional parameter tuning. While this paradigm allows for consistency in comparing the embedding spaces, it could be the case that the embeddings may perform better on certain languages, confounding the evaluation of the structural similarities between embedding spaces. For the ground-truth seed dictionary pairs, aligned seeds are provided in sets of 5,000 training pairs and 1,500 testing pairs. This benchmark was further criticized in [56] due to the large presence of proper nouns, which are considered to be only referential and contain limited lexical meaning, in Wikipedia articles, potentially over-inflating the performance metrics of systems tested on this benchmark.

一个现代的、经常被引用的 BLI 数据集工具包是脸书缪斯数据集[42].该数据集包含 30 种语言的单语嵌入和种子词典，以及 110 种语言的双语种子词典对。对于结合了单语嵌入的语言，所有这样的嵌入都是使用

FastText 算法生成的，该算法是在相应语言的维基百科副本上训练的。每组嵌入都是使用 Skip-gram 模型生成的，嵌入维数设置为 300，没有额外的参数调整。虽然这种范式允许在比较嵌入空间时保持一致性，但也可能出现这样的情况，即嵌入可能在某些语言上表现得更好，混淆了嵌入空间之间结构相似性的评估。对于地面实况种子字典对，在 5，000 个训练对和 1，500 个测试对的集合中提供对齐的种子。这一基准在[56]由于在 Wikipedia 文章中存在大量专有名词，这些专有名词被认为只是参考性的，并且包含有限的词汇意义，因此可能会过度夸大在该基准上测试的系统的性能指标。

The most popular alternative to the MUSE dataset is that compiled in [47], typically referred to as DINU. This dataset was compiled automatically from Europarl [44], a compilation of European parliament proceedings in 21 languages, typically packaged for evaluation systems into four primary languages: English, Finnish, German and Spanish [56]. Training translation pairs are split by word frequency into five buckets: 1-5K, 5-20K, 20-50K, 50-100K, 100K-200K. Within each bucket, 1,500 testing pairs are selected, as well as non-overlapping sets of training pairs in sizes of 1K, 5K, 10K and 20K.

MUSE 数据集最流行的替代品是在[47]，一般称为 DINU。该数据集是从 Europarl [44]，以 21 种语言汇编的欧洲议会会议记录，通常为评估系统打包成四种主要语言:英语、芬兰语、德语和西班牙语[56].训练翻译对按词频拆分成 5 个桶:1-5K，5-20K，20-50K，50-100K，100K-200K。在每个桶中，选择 1500 个测试对，以及大小为 1K、5K、10K 和 20K 的训练对的非重叠集合。

10.3 Knowledge Graph Entity Alignment Benchmarks
10.4 知识图实体对齐基准

Benchmarking experiments on knowledge graph entity alignment typically required two or more source knowledge graphs with a degree of known overlapping entities. One such way of generating these dataset is to use knowledge graphs describing the same set of triples in multiple languages. The WK31 datasets are an example of this paradigm, containing knowledge graphs focusing on the DBPedia person domain across English, French and German. The WK31 dataset comes in two widely utilized variants based on the number of aligned nodes, namely WK31-15k and WK31-120k. These datasets are additionally evaluated based on the number of *inter-lingual links* (ILL), where the ILL identify the same entity across two pairs of languages. The ILLs are typically used for the testing set for many evaluations and account for a small amount of the overall dataset, representing the challenge of generating a trustworthy seed dataset for supervised methods and motivating the search for many semi- or unsupervised solutions. Experiments on this dataset are reported in [60, 66, 71]. As noted in [8], there is a great deal of variance in the number of triples, entities, relations and seeds described in several publications using the WK31 dataset. Here, we provide a summary of this dataset in Table 2, based on the metrics reported in [66].

关于知识图实体对齐的基准测试实验通常需要两个或更多个具有一定程度的已知重叠实体的源知识图。生成这些数据集的一种方法是使用知识图，用多种语言描述同一组三元组。WK31 数据集是这种范式的一个例子，它包含了侧重于跨英语、法语和德语的 DBPedia 个人领域的知识图。基于对齐节点的数量，WK31 数据集有两种广泛使用的变体，即 WK31-15k 和 WK31-120k。这些数据集还基于语际链接(ILL)的数量进行评估，其中 ILL 跨两对语言识别相同的实体。ILLs 通常用于许多评估的测试集，并占整个数据集的一小部分，这代表了为监督方法生成值得信赖的种子数据集的挑战，并激发了对许多半监督或无监督解决方案的搜索。在这个数据集上的实验在[60,66,71].如[8]，在使用 WK31 数据集的几个出版物中描述的三元组、实体、关系和种子的数量有很大的差异。这里，我们在表中提供了该数据集的摘要 2，根据中报告的指标[66].

Table 2: Seed Alignments of WK31 Datasets

表 WK31 数据集的种子比对

Dataset	Aligned Entities	Aligned Relations
WK31-15k-En-De	2,070	445
WK31-15k-En-Fr	3,116	598
WK31-120k-En-De	9,680	772
WK31-120k-En-Fr	42,378	1,127
资料组	对齐的实体	结盟关系
WK31-15k-恩代	2,070	445
WK31-15k-En-Fr	3,116	598
WK31-120k-恩代	9,680	772
WK31-120k-En-Fr	42,378	1,127

Table 3: Overlapping Alignments of DFB Datasets

表 3:DFB 数据集的重叠比对

Dataset	Relations	Entities	OT	Seeds
DFB-1	1,345	14,951	0.5	5,000
DFB-1	1,345	14,951	0.5	500
DFB-1	1,345	14,951	0.1	500
资料组	关系	实体	仇恨失控	种子
DFB 一号	1,345	14,951	0.5	5,000
DFB 一号	1,345	14,951	0.5	500
DFB 一号	1,345	14,951	0.1	500

In addition to cross-lingual knowledge graph datasets, several studies have split larger graphs into smaller components, each of which has linking entities that may be used as seeds to re-unify the graph. The advantage of these datasets is that the target graph is entirely known and well defined, allowing for the number of seed entities to be scaled up and down and thus helping to experimentally validate the necessary amount of 'overlap' needed to effectively perform the alignment task. One such group of datasets named DFB-1, DFB-2 and DFB-3, constructed by [70], are constructed by randomly sampling triples from Freebase while specifying an overlapping threshold (OT). A summary of the DFB datasets is provided in 3.

除了跨语言知识图数据集之外，一些研究已经将较大的图分割成较小的组件，每个组件都具有链接实体，这些链接实体可以用作重新统一图的种子。这些数据集的优点在于，目标图是完全已知的并且被很好地定义，允许种子实体的数量被放大和缩小，从而有助于通过实验验证有效执行对齐任务所需的“重叠”的必要量。一组这样的数据集命名为 DFB 1 号、DFB 2 号和 DFB 3 号，由[70]，是通过在指定重叠阈值(OT)的同时从 Freebase 中随机采样三元组来构建的。中提供了 DFB 数据集的摘要 3.

The approach of splitting a larger graph into many subgraphs is also applied to cross-lingual knowledge graphs as well, as is the case with the DBP15k dataset [63]. This dataset uses DBPedia entities in Chinese, English, French and Japanese, creating four sets of ILLs, each containing 15,000 seed pairs, and is experimented with in [63, 69, 64, 65, 72]. A detailed description of the dataset is provided in 4.

将较大的图分割成许多子图的方法也适用于跨语言知识图，如 DBP15k 数据集的情况[63]. 该数据集使用中文、英文、法文和日文的 DBPedia 实体，创建四组 i11，每组包含 15，000 个种子对，并在[63,69,64,65,72]. 中提供了数据集的详细描述 4.

11 Summary and Open Questions

12 总结和开放式问题

To summarize, we conducted a survey of the literature on the task of aligning diverse embedding spaces output by various neural networks for creating low-dimensional representations of data. These methods have been thoroughly studied in the spaces of both natural language and graphs. The majority of these methods aim to learn alignment models between spaces of the same underlying data type, i.e. words-to-words or graphs-to-graphs, typically with the alignment meant to bridge the gap between languages. We find that there is significantly less research in bridging the gap between unlike embedding spaces, for example sentence embeddings and knowledge graphs, which we believe will provide significant gains in the fields of information extraction and data integration. By identifying this gap and outlining existing methodologies we hope this survey provides an entry point for other researchers with a shared goal of aligning embedding spaces of diverse data types.

总之，我们进行了一项文献调查，该文献的任务是排列各种神经网络输出的不同嵌入空间，以创建数据的低维表示。这些方法已经在自然语言和图形的空间中被彻底地研究过。这些方法中的大多数旨在学习相同底层数据类型的空间之间的对齐模型，即词到词或图到图，通常对齐意味着弥合语言之间的差距。我们发现，在弥合不同嵌入空间之间的差距方面的研究明显较少，例如句子嵌入和知识图，我们认为这将在信息提取和数据集成领域提供重大收益。通过确定这一差距和概述现有的方法，我们希望这项调查为其他研究人员提供一个切入点，他们的共同目标是调整不同数据类型的嵌入空间。

Table 4: Metrics of DFP Datasets

表 DFP 数据集的指标

Dataset	Language	Entities	Relations	Triples	Seeds
DBP15k-ZH-EN	Chinese	66,469	2,830	153,929	15,000
DBP15k-ZH-EN	English	98,125	2,317	237,674	15,000
DBP15k-FR-EN	French	66,858	1,379	192,191	15,000
DBP15k-FR-EN	English	105,889	2,209	278,590	15,000
DBP15k-JA-EN	Japanese	65,744	2,043	164,373	15,000
DBP15k-JA-EN	English	95,680	2,096	233,319	15,000
资料组	语言	实体	关系	增至三倍	种子
ZH 地区	中国人	66,469	2,830	153,929	15,000
ZH 地区	英语	98,125	2,317	237,674	15,000
DBP15k-FR-EN	法语	66,858	1,379	192,191	15,000
DBP15k-FR-EN	英语	105,889	2,209	278,590	15,000
DBP15k-JA-EN	日本人	65,744	2,043	164,373	15,000
DBP15k-JA-EN	英语	95,680	2,096	233,31	15,00

				9	0

References

参考

[1] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160, 2017.

[2] 亚历山大·拉特纳、斯蒂芬·h·巴赫、亨利·r·埃伦贝尔、杰森·艾伦·弗里斯、吴森和克里斯托弗·雷。浮潜:在监管薄弱的情况下快速创建训练数据。CoRR，abs/1711.10160，2017。

[3] Ann Irvine and Chris Callison-Burch. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310, jun 2017.

[4] 安·欧文和克里斯·卡利森·伯奇。双语词汇归纳综合分析。计算语言学，43(2):273–310，2017年6月。

[5] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation.

[6] 托马斯·米科洛夫、阔克·v·勒和伊利亚·苏茨基弗。利用语言间的相似性进行机器翻译。
CoRR, abs/1309.4168, 2013.
更正，abs/1309.4168，2013年。

[7] Sebastian Ruder. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902, 2017.

[8] 塞巴斯蒂安·鲁德。跨语言嵌入模型综述。更正，abs/1706.04902，2017。

[9] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *SIGMOD Rec.*, 35(3):34–41, September 2006.

[10] 崔南渊，宋一立和韩孝义。本体映射综述。西格蒙德记录。，35(3):34-41，2006年9月。

[11] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610, 2014. Evgeniy Gabrilovich Wilko Horn Ni Lao Kevin Murphy Thomas Strohmann Shaohua Sun Wei Zhang Geremy Heitz.

[12] Xin Luna Dong、Evgeniy Gabrilovich、Geremy Heitz、Wilko Horn、Ni Lao、、Thomas Strohmann、Sun和。知识库:概率知识融合的网络级方法。第20届ACM SIGKDD知识发现和数据挖掘国际会议，2014年KDD，美国纽约州，2014年8月24 - 27日，第601-610页，2014。叶夫根尼·加布利洛维奇·威尔科·霍恩倪老托马斯·斯特罗门张·格雷米·海茨。

[13] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[14] 马克西米利安·尼克尔、凯文·墨菲、福尔克·特雷普和叶夫根尼·加布利洛维奇。知识图的关系机器学习综述。IEEE会议录，104(1):11–33，2016。

[15] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs, 2020.

[16] 孙泽群、张庆恒、、陈慕豪、法拉纳兹·阿克拉米和李。基于嵌入的知识图实体对齐基准研究，2020年。

[17]Raphael Hoffmann, Luke S. Zettlemoyer, and Daniel S. Weld. Extreme extraction: Only one hour per relation.

[18]拉斐尔·霍夫曼、卢克·塞特勒莫耶和丹尼尔·威尔德。极端提取:每个关系只有一个小时。
CoRR, abs/1506.06418, 2015.
更正，abs/1506.06418，2015年。

[19] Marjorie Freedman, Lance Ramshaw, Elizabeth Boschee, Ryan Gabbard, Gary Kratkiewicz, Nicolas Ward, and Ralph Weischedel. Extreme extraction - machine reading in a week. pages 1437–1446, 01 2011.

[20] 马乔里·弗里德曼、兰斯·拉姆肖、伊丽莎白·博斯彻、瑞安·加巴德、加里·克拉特凯维奇、尼古拉斯·沃德和拉尔夫·韦斯切德尔。极端提取-机器阅读一周。第1437–1446页，2011年1月。

[21] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data.

[22] 迈克·明茨、史蒂文·比尔、里恩·斯诺和丹·茹拉夫斯基。无标记数据关系抽取的远程监控。

[23] Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2019;2018;.

[24] 亚里沙·斯米尔诺娃和菲利普·库德雷-莫鲁。使用远程监督的关系抽取:综述。ACM 计算调查(CSUR)，51(5):1–35，2019；2018;。

[25] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 73–78, New York, NY, USA, 2013. ACM.

[26] 本杰明·罗斯，塔希洛·巴斯，迈克尔·韦根和迪特里希·克拉科夫。远程监控降噪方法综述。《2013 年自动化知识库建设研讨会论文集》，AKBC '13，第 73-78 页，美国纽约州纽约市，2013 年。ACM。

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[28] 托马斯·米科洛夫，程凯，格雷戈·科拉多和杰弗里·迪恩。向量空间中单词表示的有效估计，2013。

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[30] 杰弗里·潘宁顿，理查德·索彻，克里斯托弗·曼宁。Glove:单词表示的全局向量。《2014 年自然语言处理经验方法会议论文集》(EMNLP)，第 1532-1543 页，卡塔尔多哈，2014 年 10 月。计算语言学协会。

[31] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification.

[32] 阿曼德·朱林，爱德华·格雷夫，皮奥特·博雅诺夫斯基和托马斯·米科洛夫。高效文本分类的窍门。*CoRR*, abs/1607.01759, 2016.
更正，abs/1607.01759，2016。

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[34] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。你需要的只是关注。更正，abs/1706.03762，2017。

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[36] 雅各布·德夫林、张明蔚、肯顿·李和克里斯蒂娜·图塔诺娃。BERT:用于语言理解的深度双向转换器的预训练。CoRR，abs/1810.04805，2018。

[37] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[38] 马修·e·彼得斯、马克·诺依曼、莫希特·伊耶、马特·加德纳、克里斯托弗·克拉克、肯顿·李和卢克·塞特勒莫耶。深层语境化的词语表达。CoRR，abs/1802.05365，2018。

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[40] 亚历克·拉德福德、杰夫·吴、雷文·柴尔德、大卫·栾、达里奥·阿莫代伊和伊利亚·苏茨基弗。语言模型是无人监督的多任务学习者。2019.

[41] Xunjie Zhu and Gerard de Melo. Sentence analogies: Exploring linguistic relationships and regularities in sentence embeddings, 2020.

[42] 朱勋杰和杰勒德德·梅洛。句子类比:探索句子嵌入中的语言关系和规律，2020。

[43] Nada Almarwani, Hanan Aldarmaki, and Mona Diab. Efficient sentence embedding using discrete cosine transform. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3663–3669, 2019.

[44] 纳达·阿尔马尔瓦尼、哈南·阿尔达尔马基和莫娜·迪亚卜。基于离散余弦变换的高效句子嵌入。《2019 年自然语言处理经验方法会议和第九届国际自然语言处理联合会议(EMNLP-IJCNLP)论文集》，第 3663–3669 页，2019。

[45] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors, 2015.

[46] Ryan Kiros、Yukun Zhu、Ruslan Salakhutdinov、Richard S. Zemel、Antonio Torralba、Raquel Urtasun 和 Sanja Fidler。跳过思维向量，2015。

[47] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In

[48] Lajanugen Logeswaran 和 Honglak Lee。学习句子表征的有效框架。在…里
International Conference on Learning Representations, 2018.
2018 年国际学习表征会议。

[49] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[50] Alexis Conneau、Douwe Kiela、Holger Schwenk、loc Barrault 和 Antoine Bordes。来自自然语言推理数据的通用句子表示的监督学习。《2017 年自然语言处理经验方法会议论文集》，670-680 页，丹麦哥本哈根，2017 年 9 月。计算语言学协会。

[51] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[52] 尼尔斯·雷默斯和伊琳娜·古雷维奇。句子-伯特:使用暹罗伯特网络的句子嵌入，2019。

[53] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, 2018.

[54] 米克尔·阿特克斯和霍尔格·施文克。零镜头跨语言迁移的大规模多语言句子嵌入及超越，2018。

[55] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 09 2017.

[56] 王全、毛振东、王斌和李果。知识图嵌入:方法和应用综述。IEEE 知识与数据工程汇刊，PP:1－1，09 2017。

[57] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013.

[58] Antoine Bordes、Nicolas Usunier、Alberto García-Durán、Jason Weston 和 Oksana Yakhnenko。翻译用于多关系数据建模的嵌入。《神经信息处理系统进展》26:2013 年第 27 届神经信息处理系统年会，编辑克里斯托弗·j·c·布尔吉斯、莱昂·博图、邹斌·格拉马尼和基利安·q·温伯格。2013 年 12 月 5 日至 8 日在美国内华达州太浩湖召开的会议记录，第 2787-2795 页，2013 年。

[59] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zhigang Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[60] 王震、张建文、封建林和陈志刚。基于超平面平移的知识图嵌入。2014 年在 AAAI。

[61] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press, 2015.

[62] 林、、孙茂松、、。用于知识图完成的学习实体和关系嵌入。《第二十九届 AAAI 人工智能会议论文集》，AAAI，2015 年，第 2181-2187 页。AAAI 出版社，2015。

[63] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 809–816, Madison, WI, USA, 2011. Omnipress.

[64] 马克西米连·尼克尔、沃克·特雷斯特普和汉斯·彼得·克里格尔。多关系数据集体学习的三向模型。《第 28 届国际机器学习会议论文集》，ICML，2011 年，第 809-816 页，美国威斯康星州，麦迪逊，2011 年。全媒体。

[65] Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases, 2014.

[66] 杨碧山、温头一、、高剑锋和。在知识库中嵌入用于学习和推理的实体和关系，2014。

[67] Théo Trouillon, Christopher R Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard.

Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research (JMLR)*, 18(130):1–38, 2017.

[68] Théo Trouillon、Christopher R Dance、ric Gaussier、Johannes Welbl、Sebastian Riedel 和 Guillaume Bouchard。基于复张量分解的知识图完备化。机器学习研究杂志(JMLR)，18(130):1–38，2017。

[69] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings, 2017.

[70] 蒂姆·德特默斯，帕斯夸莱·米纳维尼，庞图斯·斯坦内托普，塞巴斯蒂安·里德尔。卷积 2d 知识图嵌入，2017。

[71] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications, 2017.

[72] 林子幸·汉密尔顿、雷克斯·英和朱尔·莱斯科维奇。图形上的表征学习:方法与应用，2017。

[73] Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. Modeling relation paths for representation learning of knowledge bases, 2015.

[74] 林、、栾焕波、孙茂松、饶思伟和。知识库表征学习的关系路径建模，2015。

[75] Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space, 2015.

[76] 凯尔文·古、约翰·米勒和珀西·梁。在向量空间中遍历知识图，2015。

[77] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016.

[78] 托马斯·n·基普夫和马克斯·韦林。使用图卷积网络的半监督分类，2016。

[79] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[80] 米克尔·阿尔特塞、戈尔卡·拉巴卡和埃内科·阿吉雷。在(几乎)没有双语数据的情况下学习双语单词嵌入。《计算语言学协会第 55 届年会论文集》(第 1 卷:长篇论文)，第 451–462 页，加拿大温哥华，2017 年 7 月。计算语言学协会。

[81] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, 2018.

[82] 米克尔·阿尔特塞、戈尔卡·拉巴卡和埃内科·阿吉雷。单词嵌入的完全无监督跨语言映射的鲁棒自学习方法，2018。

[83] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[84] Alexis Conneau、Guillaume Lample、Marc'Aurelio Ranzato、Ludovic Denoyer 和 Hervé Jégou。没有平行数据的单词翻译。更正，abs/1710.04087，2017。

[85] David Alvarez-Melis and Tommi S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *CoRR*, abs/1809.00013, 2018.

[86] 大卫·阿尔瓦雷斯-梅利斯和汤米·s·雅克拉。单词嵌入空间的格罗莫夫-瓦瑟斯坦对齐。CoRR，abs/1809.00013，2018。

[87] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. 5, 11 2004.

[88] 菲利普·科恩。一个用于统计机器翻译的并行语料库。5，11 2004.

[89] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[90] 伊利亚·苏茨基弗、奥里奥尔·维尼亚尔斯和阔克诉勒。用神经网络进行序列间学习。更正，abs/1409.3215，2014 年。

[91] Natalya Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33:65–70, 12 2004.

[92] 娜塔莉亚·诺伊。语义集成:基于本体的方法综述。西格蒙德记录，33:65–70，2004 年 12 月。

[93] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem, 2014.

[94] 乔治亚娜·迪努，安吉丽·拉扎里杜和马尔科·巴罗尼。通过减轻傲慢问题改善零射击学习，2014 年。

[95] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, 2015. Association for Computational Linguistics.

[96] 赵星，，刘超和林。双语单词翻译中的规范化单词嵌入和正交变换。《计算语言学协会北美分会 2015 年会议论文集:人类语言技术》，1006-1011 页，科罗拉多州丹佛，2015 年。计算语言学协会。

[97] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. pages 2289–2294, 01 2016.

[98] 米克尔·阿尔特塞、戈尔卡·拉巴卡和埃内科·阿吉雷。学习单词嵌入的原则性双语映射，同时保持单语不变性。第 2289–2294 页，2016 年 1 月。

[99] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, February 2018.

[100] 米克尔·阿尔特塞、戈尔卡·拉巴卡和埃内科·阿吉雷。用线性变换的多步框架概括和改进双语单词嵌入映射。《第三十二届 AAAI 人工智能会议论文集》，第 5012-5019 页，2018 年 2 月。

[101] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. 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)*, pages 270–280, Beijing, China, "jul" 2015. Association for Computational Linguistics.

[102] 安吉丽卡·拉扎里杜、乔治亚娜·迪努和马尔科·巴罗尼。傲慢与污染:钻研零镜头学习的跨空间映射。《计算语言学协会第 53 届年会暨第 7 届自然语言处理国际联合会议论文集》(第 1 卷:长论文)，第 270–280 页，中国北京，“2015 年 7 月”。计算语言学协会。

[103] ChengYue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. FRAGE: frequency-agnostic word representation. *CoRR*, abs/1809.06858, 2018.

[104] 宫、狄鹤、许坦、、、。FRAGE:与频率无关的单词表示。CoRR，abs/1809.06858，2018。

[105] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. 02 2017.

[106] 齐家·穆、苏马·巴特和普拉莫德·维斯瓦纳特。除顶部之外的所有:简单有效的文字表示后处理。02 2017.

[107] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herv˜’e Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[108] Armand Joulin、Piotr Bojanowski、Tomas Mikolov、herv ’ e jéGou 和 Edouard Grave。翻译中的损失:用检索标准学习双语单词映射。《2018 年自然语言处理经验方法会议论文集》，2018。

[109] Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *CoRR*, abs/1902.00508, 2019.

[110] 戈兰·格拉瓦斯、罗伯特·李奇科、塞巴斯蒂安·鲁德和伊万·武里奇。如何(恰当地)评估跨语言单词嵌入:强基线、比较分析和一些误解。更正，abs/1902.00508，2019。

[111] Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction, 2019.

[112] Yova Kementchedjhieva、Mareike Hartmann 和 Anders sgaard。迷失在评价中:双语词典归纳的误导基准，2019。

[113] Ashwinkumar Ganesan, Frank Ferraro, and Tim Oates. Locality preserving loss to align vector spaces, 2020.

[114] 阿什温库马尔·加内桑、弗兰克·费拉罗和蒂姆·奥茨。保局损失对齐向量空间，2020。

[115] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax, 2017.

[116] 塞缪尔·史密斯、戴维·h·p·图尔班、史蒂文·汉布林和尼尔斯·汉默拉。离线双语词向量，正交变换和反向 softmax，2017。

[117] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, 2019. Association for Computational Linguistics.

[118] 巴隆·帕特拉、乔尔·鲁本·安东尼·莫尼斯、萨尔萨克·加格、马修·r·葛姆雷和格雷厄姆·纽比格。非等距嵌入空间中半监督的双语词典归纳。《计算语言学协会第 57 届年会论文集》，第 184-193 页，意大利佛罗伦萨，2019 年。计算语言学协会。

[119] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multi-lingual knowledge graph embeddings for cross-lingual knowledge alignment. *CoRR*, abs/1611.03954, 2016.

[120] 陈慕豪、田英涛、杨默涵和卡洛·扎尼奥洛。面向跨语言知识对齐的多语言知识图嵌入。更正，abs/1611.03954，2016。

[121] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. One-shot relational learning for knowledge graphs, 2018.

[122] 熊，，莫宇，，常，郭晓晓和威廉·王洋。知识图的一次性关系学习，2018。

[123] Chuxu Zhang, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V. Chawla. Few-shot knowledge graph completion, 2019.

[124] 张楚旭、姚华秀、、、和 Nitesh V. Chawla。少拍知识图补全，2019。

[125] Zequn Sun, Wei Hu, and Chengkai Li. Cross-lingual entity alignment via joint attribute-preserving embedding.

[126] 孙泽群，，，李。基于联合属性保持嵌入的跨语言实体对齐。
CoRR, abs/1708.05045, 2017.
更正，abs/1708.05045，2017。

[127] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. pages 349–357, 01 2018.

[128] 、吕、、蓝晓涵、。基于图卷积网络的跨语言知识图对齐。第 349–357 页，2018 年 1 月。

[129] Xiaofei Shi and Yanghua Xiao. Modeling multi-mapping relations for precise cross-lingual entity alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 813–822, Hong Kong, China, November 2019. Association for Computational Linguistics.

[130] 史和萧。为跨语言实体精确对齐建立多映射关系模型。《2019 年自然语言处理经验方法会议和第九届国际自然语言处理联合会议(EMNLP-IJCNLP)论文集》，第 813–822 页，中国香港，2019 年 11 月。计算语言学协会。

[131] Shengnan Li, Xin Li, Rui Ye, Mingzhong Wang, Haiping Su, and Yingzi Ou. Non-translational alignment for multi-relational networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4180–4186. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[132] 、李欣、睿烨、、苏和欧。多关系网络的非翻译比对。《第二十七届人工智能国际联合会议论文集》，IJCAI-18，第 4180-4186 页。人工智能组织国际联席会议，7 2018。

[133] Xin Li, Huiting Hong, Lin Liu, and William K. Cheung. A structural representation learning for multi-relational networks, 2018.

[134] 李欣，洪慧婷，刘林，和张广昌。多关系网络的结构表示学习，2018。

[135] Liwei Cai and William Yang Wang. KBGAN: Adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1470–1480, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[136] 蔡力伟和威廉王洋。KBGAN:知识图嵌入的对抗学习。计算语言学协会北美分会 2018 年会议论文集:人类语言技术，第 1 卷(长论文)，第 1470-1480 页，路易斯安那州新奥尔良，2018 年 6 月。计算语言学协会。

[137] Xixun Lin, Hong Yang, Jia Wu, Chuan Zhou, and Bin Wang. Guiding cross-lingual entity alignment via adversarial knowledge embedding. 11 2019.

[138] 林锡训，洪洋，，，和王斌。通过对立知识嵌入引导跨语言实体对齐。11 2019.

[139] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Iterative entity alignment via joint knowledge embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4258–4264. AAAI Press, 2017.

[140] 、谢若冰、、孙茂松。通过联合知识嵌入的迭代实体对齐。《第 26 届国际人工智能联合会议论文集》，IJCAI'17，第 4258–4264 页。AAAI 出版社，2017。

[141] Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The World Wide Web Conference*, WWW '19, page 3130–3136, New York, NY, USA, 2019. Association for Computing Machinery.

[142] 裴，陆羽，贺道夫，张。意识到程度差异的知识图嵌入半监督实体对齐。在万维网会议中，WWW '19，第 3130–3136 页，美国纽约州纽约市，2019 年。计算机协会。

[143] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. pages 4396–4402, 07 2018.

[144] 孙泽群，，张庆恒，瞿渝中。基于知识图嵌入的自举实体对齐。第 4396–4402 页，2018 年 7 月。

[145] Chaoqi Chen, Weiping Xie, Tingyang Xu, Yu Rong, Wenbing Huang, Xinghao Ding, Yue Huang, and Junzhou Huang. Unsupervised adversarial graph alignment with graph embedding, 07 2019.

[146] 、陈、、、许、俞蓉、、丁兴浩、、黄。带有图嵌入的无监督对抗图对齐，07 2019。

[147] Zuohui Fu, Yikun Xian, Shijie Geng, Yingqiang Ge, Yuting Wang, Xin Dong, Guang Wang, and Gerard de Melo. Absent: Cross-lingual sentence representation mapping with bidirectional gans, 2020.

[148] 傅、冼一坤、耿世杰、葛英强、、董鑫、和杰拉德。缺席:使用双向 gans 的跨语言句子表征映射，2020。

[149] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[150] 伊利亚·苏茨基弗、奥里奥尔·维尼亚尔斯和阔克诉勒。神经网络的序列对序列学习，2014。

[151] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[152] Myle Ott、Sergey Edunov、Alexei Baevski、Angela Fan、Sam Gross、Nathan Ng、David Grangier 和 Michael Auli。fairseq:一个快速、可扩展的序列建模工具包。在 NAACL-HLT 2019 年会议录:演示，2019 年。

[153] Channy Hong, Jaeyeon Lee, and Jungkwon Lee. Unsupervised interlingual semantic representations from sentence embeddings for zero-shot cross-lingual transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7944–7951, 04 2020.

[154] Channy Hong，Jaeyeon Lee 和 Jungkwon Lee。零触发跨语言迁移的无监督语际语义表达。AAAI 人工智能会议论文集，34:7944–7951，04 2020。

[155] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[156] 朱俊彦，朴泰星，菲利普·伊索拉和阿列克谢·埃夫罗斯。使用循环一致对抗网络的不成对图像到图像翻译。在计算机视觉(ICCV)，2017 IEEE 国际会议上，2017。