KAER：用于实体解析的知识增强预训练语言模型

Liri Fang[1*]，**Lan Li**[1*]，**Yiren Liu**[2],**Vetle I. Torvik**[1],**Bertram Lud¨ascher**[1]

[1] 伊利诺伊大学厄巴纳-香槟分校信息科学学院 [2] 伊利诺伊大学厄巴纳-香槟分校信息学 lirif2@illinois.edu、lanl2@illinois.edu、yirenl2@illinois.edu、vtorvik@illinois.edu、ludaesch@illinois.edu

arXiv:2301.04770v1 [cs.CL] 12 Jan 2023

摘要

数十年来，实体解析一直是数据清洗研究中的一项重要且得到充分研究的任务。现有研究探讨了利用预训练语言模型执行实体解析的可行性，并取得了有前景的结果。然而，很少有研究讨论通过注入领域知识来提升预训练语言模型在实体解析任务上的性能。在本研究中，我们提出**知**识**增**强**实**体**解**析（*KAER*），这是一种新颖的框架，其名称源于利用外部知识增强用于实体解析的预训练语言模型。我们讨论了利用不同知识增强和提示方法来提升实体解析性能的结果。我们的模型优于现有最先进的实体解析方法Ditto。具体而言，1）*KAER* 在"脏数据"上表现得更加稳健，并取得了更好的结果；2）借助更通用的知识注入，*KAER*在文本数据集以及在线产品领域的数据集上优于现有的基线模型；3）*KAER* 在高度特定领域的数据集（例如引文数据集）上取得了具有竞争力的结果，而这类数据集需要在未来工作中注入专家知识。

引言

近期研究表明，使用基于 *Transformer* 的预训练语言模型 （PLM）能够出色地执行各种自然语言处理任务（Min 等人，2021）。然而，鲜有研究探讨预训练语言模型在数据清洗领域的应用（Li 等，2020；Narayan 等人，2022；Vos、D¨ohmen 和 Schelter，2022）。实体解析是一项常见的数据清洗任务，旨在识别数据库内或不同数据库中指向同一现实世界实体的条目（Christen，2012）。

大多数现有的实体解析技术都假设来自不同数据源的记录具有相同的模式（Elmagarmid、Ipeirotis 和 Verykios，2007）。然而，在许多情况下，原始记录来自异构数据源，并采用不同的模式（Enr´ıquez 等人，2017；Li 等人，2021b）。此外，源数据通常来自不同的领域（例如出版物、在线产品、音乐家）并采用不同的格式（例如数值型、文本型、地点信息）。所有这些因素都增加了从业者在缺乏有关数据的领域特定信息的先验知识时执行实体解析任务的难度。因此，我们假设，增强模式级别和实体级别的外部知识可以改进实体解析任务。

随着基于 Transformer 的预训练语言模型的发展，近期研究日益关注实体解析问题（Li 等，2020；Trabelsi、Heflin 和 Cao，2022）。然而，现有研究表明，如果仅将序列化的实体对输入预训练语言模型进行分类，效果可能并不理想。Ditto（Li 等，2020）注入领域信息：预定义实体类型（即产品和数值），并在将序列化的实体对输入预训练语言模型之前规范数值格式，以提升性能。

我们进一步拓展这一思路，在模式级别和实体级别注入更多外部知识。模式级别的知识注入旨在根据数据值推断每一列的细粒度语义类型（例如 ALBUM、ARTIST、PUBLISHER）。在实体级别，我们从 WikiData 中识别实体提及，并在初始文本中标注已链接实体的语义类型信息。此外，将外部知识注入初始实体对时采用的不同格式可能会影响预训练语言模型的性能。因此，本研究进一步探索了三种提示方法：空格、斜杠和预训练语言模型的受约束调优。

总而言之，我们以最先进的方法 Ditto（Li 等，2020）为基础，提出了一个用于**知**识**增**强**实**体**消**解（KAER）的框架:

- 使用**列语义实体**类型（CST）推断和实体链接（EL），以便将领域特定信息作为附加信号注入预训练语言模型。
- 利用三种提示方法，将获取的知识更好地融入预训练语言模型。
- 分析不同的知识注入与提示方法组合在来自不同领域和数据类型的实体解析任务中的有效性。

[*]这些作者贡献相同。版权所有 © 2022，人工智能促进协会（www.aaai.org）。保留所有权利。本文已被接收，将发表于与 AAAI 2023 联合举办的自然语言处理知识增强方法研讨会。

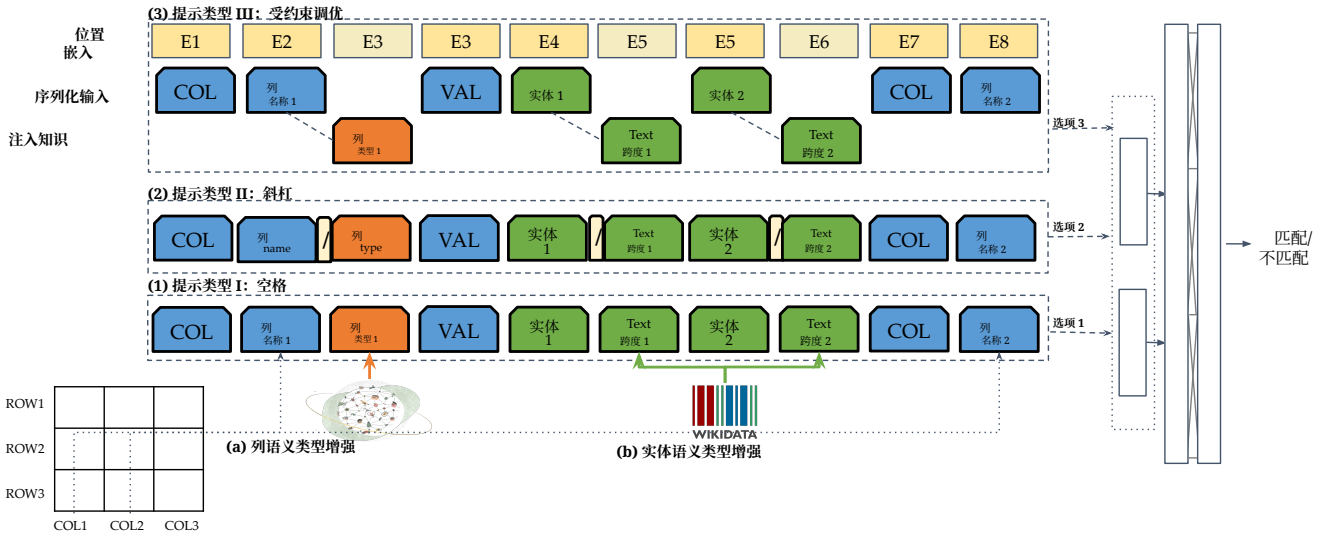

图 1：KAER 的框架。首先，采用 (a) 列语义类型增强和 (b) 实体语义类型增强，从输入表的列级别和实体级别注入领域知识。随后，在注入外部知识后提供三种提示类型选项：(1) 使用空格；(2) 使用“/”；(3) 使用受约束调优。通过知识注入和提示方法，将序列化后的序列输入预训练语言模型。

相关工作

用于实体解析的预训练模型

近期有一些研究将基于 Transformer 的预训练语言模型应用于实体解析任务。Paganelli 等人（2022年）发现，仅对 BERT 进行微调就能提升匹配/不匹配分类任务的表现，并将输入序列视为一对记录。Li 等人（2021年a）利用预训练语言模型构建孪生网络结构，以提高预训练语言模型在阻塞阶段的效率。Li 等，2020提出的 Ditto 是基于预训练语言模型（即 RoBERTa）的最先进实体匹配系统。此外，Ditto 通过注入领域知识、总结关键信息，并使用更具挑战性的示例扩充训练数据，为实体解析提供了更深入的语言理解能力。

知识增强

近期关于语义列类型增强和实体链接的研究，可以在模式级别和实体级别将外部领域知识注入预训练语言模型。语义列类型增强可以为有或没有现有列名称的列注入特定领域知识。现有研究（Hulsebos 等人，2019；Suhara 等人，2022）采用深度学习方法检测列级别的语义数据类型。实体链接（Li 等，2020）指将自然语言文本中出现的实体提及与外部知识库（如 WikiData）中的对应实体关联起来。Ayoola 等人（2022）提出了一种实体链接方法，通过在 WikiData 上微调预训练语言模型来实现实体链接；本研究采用了该方法。

“提示”是对语言模型进行条件化的方法（Liu 等人，2021年）。Kojima 等人（2022年）提出了 Zero-shot-CoT，这是一种单次零样本提示，凸显了不同提示模板会影响语言模型的性能。此外，将识别出的外部知识注入预训练语言模型的方法也至关重要。

Liu 等人（2020年）提出了 K-BERT 模型，通过在增强后的输入序列中添加软位置编码和可见矩阵，减少注入原始输入及相应模型中的知识噪声。

方法论

实体解析任务的符号表示

本文对实体解析任务的形式化定义基于 Li 等人（2020年）的相关工作。实体解析任务的输入由一组 $M \subseteq D_1 \times D_2$ 构成，其中 D_1 和 D_2 是两个包含重复条目的数据条目集合。对于每个数据条目，$e \in (col_i, val_i)_{1 \le i \le N}$，其中 N 是列数，而 $(e_1, e_2) \in M$。本文讨论的任务是：对于每一对数据条目 $(e_1, e_2) \in M$，判断 e_1 和 e_2 是否指向同一个数据实体。

模型概述

本研究提出了一个用于实体解析的新框架：KAER（见图 1）。KAER 使用预训练语言模型进行实体解析，并包含三个用于知识增强的模块：a) 列语义类型增强，b) 实体语义类型增强，以及 c) 三种提示类型选项。以下各节将首先介绍用于实体解析的预训练语言模型，然后介绍各个模块。

用于实体解析的预训练语言模型。 遵循 Li 等人（2020年）的工作，KAER 使用 RoBERTa 作为骨干模型。对于每对数据条目 (e_1, e_2)，将 e_1 和 e_2 的列名和值的文本上下文序列化并拼接，作为预训练语言模型的输入。使用 [CLS] 词元位置来判断 e_1 和 e_2 是否指向同一实体。用于优化分类目标的损失为：

$$\ell = -log\, p(y|s(e_1, e_2)) \quad (1)$$

其中，y 表示 e_1 和 e_2 是否指向同一实体，$s(\cdot, \cdot)$ 表示通过知识注入和提示方法对实体对进行序列化和转换。

$$s(e_i, e_j) ::= \text{[CLS]}\ serialize(e_i)\ \text{[SEP]}\ serialize(e_j)\ \text{[SEP]} \quad (2)$$

其中，$serialize(\cdot)$ 对每个数据条目进行序列化。

$$\begin{aligned} serialize(e_i) ::= &\text{[COL]}\ f(col_1, pt)\ \text{[VAL]}\ g(val_1, pt)\ ... \\ &\text{[COL]}\ f(col_N, pt)\ \text{[VAL]}\ g(val_N, pt) \end{aligned} \quad (3)$$

其中，$f(col_i, pt)$ 表示采用提示方法 pt 的列语义类型注入，而 $g(val_i, pt)$ 表示采用提示方法 pt 的实体链接注入。接下来的章节将介绍列语义类型注入、实体链接和三种提示方法。

列语义类型增强。列类型预测旨在通过考虑列内上下文来预测每一列的语义类型，从而为框架增强细粒度的语义信息。例如，作为列名的“名称”在注入“歌曲”这一语义类型后，可以提供更多语义信息。此处，KAER 使用 Sherlock（Hulsebos 等人，2019）进行列语义类型预测。具体而言，*Sherlock* 是一个多输入深度神经网络，它通过将 DBpedia 中的 78 种语义数据类型与列标题进行匹配，在 VizNet 语料库中检索到的 686,765 个数据列上进行预训练（Hulsebos 等人，2019）。

实体语义类型增强。实体语义类型增强利用实体链接方法，从给定的知识库（KB）中识别目标文本序列中的所有实体提及。具体而言，1）识别每个实体提及的文本跨度 $m \in M$，并且 2）将每个提及映射到外部知识库 $\mathcal{M}$ 中的实体集合：$M \rightarrow E$。因此，利用知识库中的额外知识来增强输入文本记录。本研究使用 WikiData 作为外部知识库，因为它涵盖广泛的领域，适用于不同数据集上的实体解析。KAER 使用 Ayoola 等人（2022年）提出的最先进方法，该方法采用 RoBERTa 模型，在实体类型判定和实体描述建模目标上进行联合预训练，以注入粗粒度实体类型。

提示方法

基于模板的提示。获取领域知识后，我们尝试使用基于文本的模板，将这些知识与初始文本输入相结合。我们使用两种不同的字符来连接原始实体提及/列名，即斜杠（“/”）和空格。之所以选择斜杠符号，是因为在一般网络文本中，它具有与“或”相当的语义。

受约束调优提示 受 Han 等人（2021年）提出的硬提示与软提示混合调优的启发，受约束调优提示通过将增强知识的可见范围限制在其对应的实体提及或列名上，来生成词元嵌入。这种领域知识注入会从两个方面修改初始输入序列，包括词元位置和词元语义。更具体地说，来自初始输入序列的词元在与基于模板的提示拼接后，其绝对位置会发生变化。此外，针对特定实体提及注入的知识会与同一序列中的周围词元共享。这两种修改可能会给框架引入错误知识并造成负担。

为限制错误知识和过载，我们借鉴 KBert（Liu 等，2020），采用受约束调优，即软位置编码和可见矩阵，作为骨干模型的额外输入。这里，知识注入序列被构造成树形结构，其中注入的知识子序列被视为分支。如图 1 所示，软位置编码使初始词元在知识注入序列中的位置编码与其在初始输入序列中的位置编码相同，并为注入的知识词元分配分支中的位置。可见矩阵 $V \in \mathcal{R}^{N \times N}$ 是一个二值矩阵，其中注入的知识词元仅对自身以及相应的实体提及或列名词元可见，并将其设为 1，如下所示：

$$V_{ij} = \begin{cases} 1, & \text{if } token_i \text{ and } token_j \text{ co-occur.} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

其中，$token_i$ 和 $token_j$ 的共现表示这些词元在初始输入序列中共现，或在领域知识注入中分别作为头部和尾部共现。这里，头部指实体提及/列名，尾部指注入的知识文本。

实验

数据集

KAER 在来自多个领域的 Magellan 数据集（Das 等，2022年）上进行了评估。本研究所用数据集的概览见表 2。Magellan 数据集包含三种类型：脏数据集、结构化数据集和文本数据集。脏数据集由结构化数据集生成，方法是随机删除属性值，并将初始值追加到另一个随机选定的属性中（Li 等，2020年）。

结果

实验结果见表 1。

知识增强在较小的数据集上表现更好。根据实验结果，融入外部知识可以提升实体解析任务的性能。尤其是在数据稀缺情境下，所提出的知识增强方法优于两种基线模型。例如，**KAER（RoBERTa+实体链接+/）** 在 Dirty/iTunes-Amazon 数据集上的表现优于基线模型 RoBERTa 和 Ditto，分别高出 9.09% 和 21.44% ↑，但差异在统计上不显著。同时，在 Structured/iTunes-Amazon 数据集上，其表现分别显著优于两种基线模型 24.45% 和 39.15% ↑。两种基线模型在较小的数据集上表现较差，因为微调时训练实例较少，语言模型可能无法学习到足够的信息来区分不同实体。使用提示

模型	脏		结构化				文本
	DBLP - Google Scholar	iTunes - 亚马逊	iTunes - 亚马逊	亚马逊 - 谷歌	DBLP - ACM	DBLP - Google Scholar	Abt - Buy
RoBERTa	95.81	72.73	61.76	73.12	98.77	95.32	88.41
Ditto	95.54	60.38	47.06	74.44	98.55	95.85	88.89
KAER（RoBERTa + CST）	95.47	62.96	80.70**++	68.70	98.77	95.05	92.27**++
KAER（RoBERTa + CST + /）	95.74	54.90*	**89.66**++**	74.85	**98.99**	95.14	**92.61**++**
KAER（RoBERTa + CST + PCT）	95.20	72.00	57.14	**76.25**	98.09*	95.62	78.24**++
KAER（RoBERTa + 实体链接）	95.12	50.00*	61.54	73.77	98.54	95.72	89.11
KAER（RoBERTa + 实体链接 + /）	95.75	**81.82**	86.21**++	74.02	98.43	**96.11***	88.19
KAER（RoBERTa + 实体链接 + PCT）	94.80**+	77.78+	67.69	75.63	97.04**+	95.29	73.80**++
KAER（RoBERTa + CST + 实体链接）	**95.87**	47.06*	59.26	72.28	98.43	95.65	89.66
KAER（RoBERTa + CST + 实体链接 + /）	95.63	63.33*	57.14+	75.15	98.34	95.67	89.81
KAER（RoBERTa + CST + 实体链接 + PCT）	94.24**++	71.11+	61.29	71.37*	96.26**++	94.93+	67.20**++

表 1：使用不同知识注入方法的实验结果，以 F1分数衡量。两种基线模型：RoBERTa（不进行领域知识注入）；以及 Ditto（RoBERTa 和 Ditto 的通用领域知识注入方法）。其他模型表示将各种知识注入和提示方法与 RoBERTa 相结合的 KAER。“+CST”表示**列语义类型**注入。“+实体链接”表示通过**实体链接**进行知识注入。“+/”表示使用斜杠进行提示。“+PCT”表示**提示约束调优**，即软位置和可见矩阵。配对T检验用于比较 KAER 注入方法与两种基线模型，即 RoBERTa(*) 和 Ditto(+)。** 和 ++表示 99% 置信水平，* 和 + 表示 95% 置信水平。

数据类型	数据集	领域	Size	# 正例	# 属性
结构化	亚马逊-谷歌	软件	11,460	1,167	3
	iTunes-亚马逊	音乐	539	132	8
	DBLP-ACM	引文	12,363	2,220	4
	DBLP-Google Scholar	引文	28,707	5,347	4
脏	iTunes-Amazon	音乐	539	132	8
	DBLP-ACM	引文	12,363	2,220	4
	DBLP-Google Scholar	引文	28,707	5,347	4
文本	Abt-Buy	产品	9,575	1,028	3

表 2：数据集概述

PCT 在文本长度较短的数据集上表现更好。例如，在序列化数据条目长度最短的 Structured/Amazon-Google 数据集上，**KAER（RoBERTa+CST+PCT）**取得了最高的 F1分数（76.25%），但差异不具统计显著性。此外，在 Structured/Amazon-Google 上，**KAER（RoBERTa+CST+PCT）**与两个基线模型相比差异显著，置信度为 99%。

知识增强在某些领域的数据集上表现不佳。 我们实验中使用的数据集涵盖了多个不同领域。采用我们提出的知识增强方法（CST 和 EL）后，音乐和在线产品等某些领域的记录能够受益于更准确的检索知识。例如，**KAER(RoBERTa+CST+/)** 在在线产品领域的文本数据集 Abt-Buy 上取得了最佳 F1 分数，且具有统计显著性，即 92.61%。**KAER (RoBERTa+EL+/)** 在音乐数据集 Dirty/iTunes-Amazon 上取得了最佳 F1 分数（81.82%），但不具有统计显著性。出现这一显著提升的一个潜在原因是，这些领域数据集中提及的实体更为通用。然而，引用等领域的出版物数据集需要的是更具领域针对性的知识，甚至是专家知识，而非通用常识。我们的知识注入方法主要用于检索通用知识。因此，对于 Structured/DBLP-ACM 这类数据集，**KAER(RoBERTa+CST)** 表现与 RoBERTa 一样好，并且与 Ditto 相比，在 95% 的置信水平下差异显著；而 **KAER(RoBERTa+CST+/)** 取得了最佳 F1 分数，为 98.99%，但不具有统计显著性。不过，其他所有 KAER 模型的表现都较差。为了进一步改进，应将引用领域的更多外部专家知识纳入知识注入。

知识增强会受到数据质量的影响。采用特定知识增强方法的 KAER 在脏数据集上的表现优于基线模型。例如，**KAER（RoBERTa+CST+实体链接）**在 Dirty/DBLP-GoogleScholar 上略优于 RoBERTa 和 Ditto。然而，在 Dirty/DBLP-GoogleScholar 上，其他所有注入方法的 F1分数均不及基线模型（RoBERTa 和 Ditto）的结果。一个可能的原因是，CST 基于脏输入产生的误导性预测会造成语义噪声，并将其传播到预训练语言模型中。另一方面，采用 PCT 进行知识增强可以提升模型处理低质量注入的鲁棒性。例如，**KAER（RoBERTa+CST+PCT）**的 F1分数高于 **KAER（RoBERTa+CST）**（72.00% 对比 62.96%）。**KAER（RoBERTa+实体链接+PCT）**与**KAER（RoBERTa+实体链接）**相比，以及**KAER（RoBERTa+CST+实体链接+PCT）**与**KAER（RoBERTa+CST+实体链接）**相比，也有类似的提升。

结论

在本研究中，我们提出了一种名为 KAER 的新型框架，通过提示技术为预训练语言模型注入外部知识，以解决实体解析问题。我们在来自不同领域和数据类型（即脏数据、结构化数据和文本数据）的一系列数据集上评估了该框架，结果表明，在某些应用中，其性能优于基线模型。进一步分析揭示了不同领域知识注入方式和提示方法在来自不同领域及数据类型的各类数据集上的有效性。结果显示，通过向预训练语言模型注入知识来提升实体解析性能，是一个很有前景的方向。未来的研究可以通过引入更多特定领域知识来改进我们的方法。

致谢

实验在 NCSA Delta 超级计算系统上进行，该系统由美国国家科学基金会通过第 OAC-2005572 号资助项目提供支持。本研究由美国国立卫生研究院（NIH）下属美国国家老龄研究所资助（项目编号：P01AG039347）。资助方未参与研究设计、数据收集与分析、发表决定或稿件撰写。

参考文献

Ayoola, T.; Tyagi, S.; Fisher, J.; Christodoulopoulos, C.; and Pierleoni, A. 2022年。ReFinED：一种高效、具备零样本能力的端到端实体链接方法。载于 *2022*年北美计算语言学协会会议论文集：人类语言技术：产业赛道，*NAACL 2022*年，209–220。

Christen, P. 2012年。数据匹配流程。载于数据匹配：记录链接、实体解析和重复项检测的概念与技术，《以数据为中心的系统与应用》，23–35。柏林，海德堡：施普林格。ISBN978-3-642-31164-2。

Das, S.；Doan, A.；G. C., P. S.；Gokhale, C.；Konda, P.；Govind, Y.；以及 Paulsen, D. 2022年。《Magellan 数据仓库》。https://sites.google.com/site/anhaidgroup/projects/data。Elmagarmid, A. K.；Ipeirotis, P. G.；以及 Verykios, V. S. 2007年。重复记录检测：综述。*IEEE* 知识与数据工程汇刊，19(1)：1–16。会议名称：IEEE 知识与数据工程汇刊。Enr´ıquez, J.；Dom´ınguez-Mayo, F.；Escalona, M.；Ross, M.；以及 Staples, G. 2017年。大数据源中的实体对账：一项系统性映射研究。《专家系统与应用》，80：14–27。Han, X.；Zhao, W.；Ding, N.；Liu, Z.；以及 Sun, M. 2021年。PTR：用于文本分类的提示调优。*CoRR*，abs/2105.11259。Hulsebos, M.；Hu, K. Z.；Bakker, M. A.；Zgraggen, E.；Satyanarayan, A.；Kraska, T.；Demiralp, C.。；以及 Hidalgo, C. A. 2019年。Sherlock：一种用于语义数据类型检测的深度学习方法。载于第*25*届 *ACM SIGKDD* 国际 *KDD* 会议论文集（*KDD 2019*），1500–1508。ACM。Kojima, T.；Gu, S. S.；Reid, M.；Matsuo, Y.；以及 Iwasawa, Y. 2022年。大语言模型是零样本推理器。*CoRR*，abs/2205.11916。Li, B.；Miao, Y.；Wang, Y.；Sun, Y.；以及 Wang, W. 2021a。提升基于 BERT 的实体解析的效率与有效性。*AAAI*人工智能大会论文集，35(15)：13226–13233。编号：15。Li, X.；Talburt, J. R.；Li, T.；以及 Liu, X. 2021b。当实体解析遇上深度学习时，相似度度量是否仍有必要？载于人工智能与应用认知计算进展，127–140。施普林格。Li, Y.；Li, J.；Suhara, Y.；Doan, A.；以及 Tan, W.-C. 2020年。利用预训练语言模型进行深度实体匹配。《*VLDB*基金会论文集》，14(1)：50–60。

Liu, P.；Yuan, W.；Fu, J.；Jiang, Z.；Hayashi, H.；以及 Neubig, G. 2021年。预训练、提示与预测：自然语言处理中的提示方法系统性综述。*CoRR*，abs/2107.13586。

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020年。K-BERT：利用知识图谱实现语言表示。载于*AAAI*会议论文集，第34卷，2901–2908。

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2021年。通过大型预训练语言模型实现自然语言处理的最新进展：一项综述。*CoRR*，a bs/2111.01243。Narayan, A.; Chami, I.; Orr, L. J.; and R´e, C. 2022年。基础模型能否帮你整理数据？*CoRR*，abs/2205.09911。Paganelli, M.; Buono, F. D.; Baraldi, A.; and Guerra, F. 2022年。分析 howBERT 如何执行实体匹配。*VLDB*基金会论文集，15(8)：1726–1738。Suhara, Y.; Li, J.; Li, Y.; Zhang, D.; Demiralp, c.; Chen, C.; and Tan, W.-C. 2022年。使用预训练语言模型标注列。载于*2022*年数据管理国际会议论文集，SIGMOD '22，1493–1503。计算机协会。

Trabelsi, M.; Heflin, J.; 和 Cao, J. 2022年。DAME：用于实体匹配的领域自适应。载于第十五届 *ACM* 国际网络搜索与数据挖掘会议论文集，1016–1024。线上活动，美国亚利桑那州：ACM。ISBN 978-1-4503-9132-0。

Vos, D.; D¨ohmen, T.; 和 Schelter, S. 2022年。通过前缀调优实现数据整理任务的参数高效自动化。载于 *NeurIPS 2022* 第一届表格表示研讨会。

附录

A. 超参数设置。基线模型和 KAER 模型采用相同的超参数设置：批量大小为 64，最大长度为 512，学习率为 3e-5，训练轮数为 20。

B. 基线模型。本文包含两个基线模型，即未进行任何知识注入的 RoBERTa，以及采用默认知识注入方法的 Ditto。

数据集	注入
脏数据/DBLP-Google	Ditto +通用
Dirty/iTunes-Amazon	Ditto +产品
Structured/iTunes-Amazon	Ditto +产品
结构化/Amazon-Google	Ditto +产品
结构化/DBLP-ACM	Ditto +通用
结构化/DBLP-Google	Ditto +通用
文本/Abt-Buy	Ditto +产品

表 3：数据集及其对应的 Ditto 默认知识注入方法

具体而言，表 3 展示了 Ditto 使用的数据集及相应的注入方法。“+ General”表示 Ditto 将七种实体类型注入相应的数据集。这七种类型包括“PERSON”“ORG”“LOC”“PRODUCT”“DATE”“QUANTITY”和“TIME”。“+ Product”表示 Ditto 将以下类型中的所有实体提及标注为“PORODUCT”：即“NORP”“GPE”“LOC”“PERSON”和“PRODUCT”。

C. 双侧配对T检验。我们采用双侧配对T检验，检验不同模型的预测结果是否存在显著差异。我们首先比较预测标签是否与真实标签相同。如果预测结果与真实标签相同，则值为 1，否则为 0。然后进行配对T检验。