Data Science and Engineering (2023) 8:124–134
https://doi.org/10.1007/s41019-023-00208-9

RESEARCH PAPER

Probing the Impacts of Visual Context in Multimodal Entity Alignment

Meng Wang[1] · Yinghui Shi[2] · Han Yang[3] · Ziheng Zhang[4] · Zhenxi Lin[4] · Yefeng Zheng[4]

Received: 7 December 2022 / Revised: 8 March 2023 / Accepted: 19 March 2023 / Published online: 4 April 2023

Abstract
We study the problem of multimodal embedding-based entity alignment (EA) between different knowledge graphs. Recent works have attempted to incorporate images (visual context) to address EA in a multimodal view. While the benefits of multimodal information have been observed, its negative impacts are non-negligible as injecting images without constraints brings much noise. It also remains unknown under what circumstances or to what extent visual context is truly helpful to the task. In this work, we propose to learn entity representations from graph structures and visual context, and combine feature similarities to find alignments at the output level. On top of this, we explore a mechanism which utilizes classification techniques and entity types to remove potentially un-helpful images (visual noises) during alignment learning and inference. We conduct extensive experiments to examine this mechanism and provide thorough analysis about impacts of the visual modality on EA.

Keywords Entity alignment · Multimodality · Visual context · Knowledge graph

1 Introduction

Entity alignment (EA) is a task aiming to find entities from different knowledge graphs (KGs) that refer to the same real-world object. It plays an important role in KG construction and knowledge fusion as KGs are often independently created and suffer from incompleteness. Most existing models for EA leverage graph structures and/or side information of entities such as name and attributes along with KG embedding techniques to achieve alignment [1, 2]. Several recent methods enrich entity representations by incorporating images, a natural component of entity profiles in many KGs such as DBpedia [3] and Wikidata [4], to address EA in a multimodal view [5–7].

While experimental results have demonstrated that incorporating visual context benefits the EA task [5, 7], it is worth noting that the use of entity images may introduce noises. An error analysis in EVA [7] pointed out that hundreds of source entities were correctly matched to their counterparts before injecting images but were mismatched with images present. Different visual representations of equivalent entities could be potential noises that induce mismatches, and there are various reasons for the visual inconsistency between two equivalent entities. One major reason is that entities naturally have multiple visual representations. As shown in Fig. 1, images (visual context) at left are dissimilar from their counterparts at right, yet they refer to same real-world entities. In addition, the incompleteness of visual data is also a challenging issue for multimodal EA, as reported in [7] that ca. 15–50% entities in the most commonly used benchmark DBP15K [8] are not provided with images.

✉ Yinghui Shi
shiyinghui@seu.edu.cn

Meng Wang
wangmengsd@outlook.com

Han Yang
han.yang6@zeekrlife.com

Ziheng Zhang
zihengzhang@tencent.com

Zhenxi Lin
chalerislin@tencent.com

Yefeng Zheng
yefengzheng@tencent.com

[1] College of Design and Innovation, Tongji University, Shanghai, China

[2] School of Cyber Science and Engineering, Southeast University, Nanjing, China

[3] ZEEKR Intelligent Technology Holding Ltd., Shanghai, China

[4] Tencent Jarvis Lab, Shenzhen, China