DiaKG: an Annotated Diabetes Dataset for Medical Knowledge Graph Construction

Dejie Chang[1], Mosha Chen[2], Chaozhen Liu[1], Liping Liu[1], Dongdong Li[1], Wei Li[1], Fei Kong[1], Bangchang Liu[1], Xiaobin Luo[1], Ji Qi[3], Qiao Jin[3], and Bin Xu[3]

[1] Miao Health {changdejie,liuchaozhen,liuliping,lidongdong,liwei,kongfei, liubangchang,luoxiaobin}@miao.cn
[2] Alibaba Group chenmosha.cms@alibaba-inc.com
[3] Tsinghua university
{jqa14,qj20}@mails.tsinghua.edu.cn,xubin@tsinghua.edu.cn

Abstract. Knowledge Graph has been proven effective in modeling structured information and conceptual knowledge, especially in the medical domain. However, the lack of high-quality annotated corpora remains a crucial problem for advancing the research and applications on this task. In order to accelerate the research for domain-specific knowledge graphs in the medical domain, we introduce DiaKG, a high-quality Chinese dataset for Diabetes knowledge graph, which contains 22,050 entities and 6,890 relations in total. We implement recent typical methods for Named Entity Recognition and Relation Extraction as a benchmark to evaluate the proposed dataset thoroughly. Empirical results show that the DiaKG is challenging for most existing methods and further analysis is conducted to discuss future research direction for improvements. We hope the release of this dataset can assist the construction of diabetes knowledge graphs and facilitate AI-based applications.

Keywords: Diabetes · Dataset · Knowledge graph.

1 Introduction

Diabetes is a chronic metabolic disease characterized by high blood glucose level. Untreated or uncontrolled diabetes can cause a range of complications, including acute ones like diabetic ketoacidosis and chronic ones such as cardiovascular diseases and diabetic nephropathy. With the rapid economic developments and changes in lifestyle, China has become the country with the most diabetes patients in the world: the prevalence of diabetes in Chinese adults is about 11.2% and still increasing[1]. The medical expenses from diabetes without complications already account for 8.5% of national health expenditure in China[2]. Cardiovascular diseases, one of the complications of diabetes, are the leading cause of death in China. Diabetic nephropathy, another diabetes complication, could "waste the wealth that we've accumulated over the past 30 years in the drains of dialysis machines" according to [3]. As a result, diabetes is a serious public health problem in the realization of "Healthy China 2030" that requires interdisciplinary innovations to solve.

Knowledge Graph (KG) has been proven effective in modeling structured information and conceptual knowledge, especially in the medical domain[4]. Medical knowledge graph is attracting attention from both academic and healthcare industries due to its power in intelligent healthcare applications, such as clinical decision support systems (CDSSs) for diagnosis and treatment[5,6], self-diagnosis utilities to assist patient evaluating health conditions based on symptoms[7]. High-quality entity and relation corpus is crucial for constructing knowledge base, however, there is no dataset dedicated to the diabetes disease at the moment. To address this issue, we introduce DiaKG, a high-quality Chinese dataset for Diabetes knowledge graph construction.

The contributions of this work are as follows:

1. To the best of our knowledge, this is the first diabetes dataset for medical knowledge graph construction at home and abroad.

2. In addition to the medical experts, we introduce AI experts to participate in the annotation process to provide the perspective of AI models, which improves the usability of the annotation data and finally benefits the end-to-end performance.

We hope the release of this corpus can help researchers develop knowledge bases for clinical diagnosis, drug recommendation, and auxiliary diagnostics to further explore the mysteries of diabetes. The datasets are publicly available at https://tianchi.aliyun.com/dataset/dataDetail?dataId=88836

2 DiaKG Construction

2.1 Data Resource

The dataset is derived from 41 diabetes guidelines and consensus, which are from authoritative Chinese journals covering the most extensive fields of research content and hotspot in recent years, including basic research, clinical research, drug usage, clinical cases, diagnosis and treatment methods, etc. Hence it is a quality-assured resource for constructing a diabetes knowledge base.

2.2 Annotation Guide

Two seasoned endocrinologists designed the annotation guide. The guide focuses on entities and relations since these two types are the fundamental elements of a knowledge graph.

Entity 18 types of entities are defined(Table.1). Nested entities are allowed; for example, ’2型糖尿病’ is a ’Disease’ entity, and ’2型’ is a ’Class’ one. Besides the nested structure, entities in DiaKG has two characteristics that stand out: 1. Entities may attribute to different types according to the contextual content, a tough task for NER models; for example, ’糖尿病’ in sentence ’糖尿病患者需控制饮食’ is a ’Disease’ type, while in the sentence ’糖尿病所致肾损伤占1/3’ serves as a ’Reason’ type; 2. Some entity types are of long span like ’Pathogenesis’ are usually consisted of a sentence.

Table 1. List of entities annotation guide

entity name	label	example
疾病	Disease	表明运动对**1型糖尿病微血管病变**的预后无改善作用
疾病的分期分型	Class	纽约心脏学会(NYHA)心功能**Ⅲ-IV级**、终末期肾病
病因	Reason	若**体重增加**，可能加重**胰岛素抵抗**
发病机制	Pathogenesis	多数患者的**β细胞完全破坏**，**胰岛水平极低**
临床表现	Symptom	已发生明确的**足趾、足掌、肢体坏疽创面**
检查方法	Test	速食面进行**混合餐耐量试验(MMTT)**
检查指标	Test_Items	快速血糖仪测量**指血(毛细血管血)血糖**
检查指标值	Test_Value	有低血糖症状并且血糖**$< 3.3mmol/L$**
药物名称	Drug	包括**COX-2抑制剂**
用药频率	Frequency	12h后按照0.5mg，**1～3次／d**
用药剂量	Amount	可根据**0.3～0.5单位/千克体重**来估算起始胰岛素总量
用药方法	Method	短效胰岛素一般在**餐前15～30min皮下注射**
非药治疗	Treatment	**认知-行为及心理干预**是通过调整患者的生活环境
手术	Operation	进行**胰岛细胞移植手术**来改善患者的胰岛情况
不良反应	ADE	贝特类可使**胆结石的发生率升高**
部位	Anatomy	糖尿病相关**微血管**和**大血管**并发症等方面的证据
程度	Level	但对于**中到重度**肾功能不全的患者需要减少剂量
持续时间	Duration	预防治疗维持**3～6个月**

Relation Relations are centered on 'Disease' and 'Drug' types, where a total of 15 relations are defined(Table.2). Relations are annotated on the paragraph level, so entities from different sentences may form a relation, which has raised the difficulty for the relation extraction task. Relation entities in the same sentence only account for 43.4% in DiaKG. The proportions of entities located in two consecutive sentences and in non-consecutive ones are 24.0% and 22.%, respectively.

2.3 The Annotation Process

The annotated process is shown in Fig.1. The process can be divided into two steps:

OCR Process The PDF files are transformed to plain text format via the OCR tool[1], where non-text data like figures and tables are manually removed. Additionally 2 annotators manually check the OCR results character by character to avoid misrecognitions, for example, 'β细胞' may be recognized as 'B细胞'.

Annotation Process 6 M.D. candidates were employed for the annotation task and were trained thoroughly by our medical experts to have a comprehensive understanding of the annotation guide. During the **trail annotation** step, we creatively invited 2 AI experts to label the data simultaneously, based on the

[1] https://duguang.aliyun.com/

Table 2. List of selected relations annotation guide

relation	example
Test_Items_Disease	血浆**酮体**增加或**酮血症**倾向往往低于正常人
Treatment_Disease	积极进行**糖尿病**防治知识的学习和宣教，**增加运动**
Class_Disease	分级**l-II级**的**充血性心力衰竭**的患者中的治疗经验有限
Anatomy_Disease	糖尿病合开各种严重慢性开发症如各种**神经病变**、视网膜病变、...等
Drug_Disease	心血管保护作用：**二甲双胍**通过有效改善**糖尿病**和非糖尿病患者的IR。
Reason_Disease	**慢性梗阻**可引起**肾积水**和肾实质萎缩，甚至发展为终末期肾病
Symptom_Disease	此分级方法...对**糖尿病**足溃疡及...更好地体现了**创面感染**和缺血的情况
Operation_Disease	接受肥胖与**糖尿病**外科手术患者...对接受**减重代谢手术**的病人
Test_Disease	5项检查(...压力觉、**温度觉**)等方法半定量评估患者的**神经病变**程度
Pathogenesis_Disease	二甲双胍可改善**IR**...具有更全面针对**T2DM**的病理生理缺陷的特点
ADE_Drug	正确使用**磺脲类药物**单药...，**轻、中度低血糖**发生率为...
Amount_Drug	**二甲双胍**(**1000mg/d**)起始治疗，其胃肠道反应发生率为24%
Method_Drug	**短效胰岛素**一般在**餐前15～30min皮下注射**
Frequency_Drug	**每日1次**基础**胰岛素**或...作为胰岛素起始治疗方案
Duration_Drug	**持续**静脉泵注**胰岛素**有利于减少血糖波动

assumption that they provided an AI model's perspective. For example, medical experts are inclined to label '成年型糖尿病(maturity-onset diabetes of the young，MODY)' as a whole entity, while AI experts regard separate annotations as '成年型糖尿病', 'maturity-onset diabetes of the young' and 'MODY' are more model-friendly. Feedback from AI experts and issues raised by the annotators were sent back to the medical expert to refine the annotation guideline iteratively. The **formal annotation** step started by the 6 M.D. candidates and 1 medical experts would give timely help when needed. **The Quility Control (QC)** step was conducted by the medical experts to guarantee the data quality, and common annotation problems were corrected in a batch mode.

The quality is evaluated by another medical expert via random sampling of 300 records. The accuracy rates of entity and relation are 90.4% and 96.5%, respectively, demonstrating the high-quality of DiaKG. The examined dataset contains 22,050 entities and 6,890 relations, which is empirically adequate for a specified disease.

3 Experiments

We conduct Named Entity Recognition(NER) and Relation Extraction(RE) experiments to evaluate DiaKG. The codebase is public on github[1].

3.1 Named Entity Recognition (NER)

We only report results from X Li et al.(2019)[8] since it is the SOTA model for NER with nested settings at the time of this writting.

[1] https://github.com/changdejie/diaKG-code

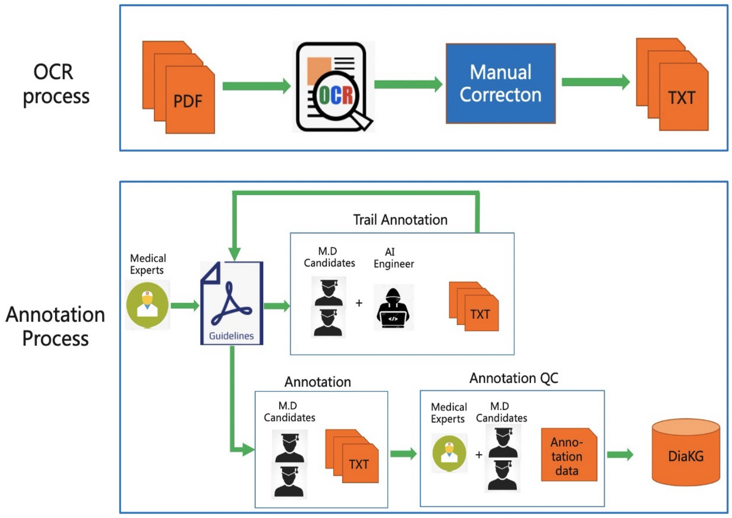

Fig. 1. The annotated process of the diabetes dataset.

3.2 Relation Extraction (RE)

The RE task is defined as giving two entities that are annotated as relation, to classify the relation type. Due to the simplified setting, we only report results from bi-directional GRU-attention[9] in this paper.

Table 3. selected NER results

Entity	precision	recall	F1
total	0.814	0.853	0.833
Drug	0.881	0.902	0.892
Disease	0.794	0.91	0.848
Pathogenesis	0.595	0.667	0.629
Symptom	0.535	0.535	0.535
Reason	0.333	0.3	0.316

Table 4. selected RE results

Relation	precision	recall	F1
total	0.839	0.837	0.836
Class_Disease	0.968	0.874	0.918
ADE_Drug	0.892	0.892	0.892
Test_Disease	0.648	0.636	0.642
Pathogenesis_Disease	0.486	0.692	0.571
Operation_Disese	0.6	0.231	0.333

4 Analysis

The experimental results are shown in Table.3 and Table.4. We only report the total result of all the entity/relation types, plus the top 2 and last 3 types' results for each task due to space limitation.

The **overall** macro-average scores for the two tasks are 83.3% and 83.5%, respectively, which are satisfying considering the multifarious types we define, also demonstrating DiaKG's high quality. For the **NER task**, the results of 'Disease' and 'Drug' types are as expected because entities of these two types exist frequently among the documents, thus leading to a higher score. The average entity length for 'Pathogenesis' type is 10.3, showing that the SOTA MRC-Bert model still can not handle the long spans perfectly; We analyzed errors of the

'Symptom' and 'Reason' types and found that the model is prone to classify entities as other types, mainly contributing to the characteristic that entity may be of different types due to the contextual content. For the **RE task**, the case study shows that entities with long distance are difficult to classify. For example, entities with 'Drug_Diesease' type usually exist in the same sub-sentence, whereas the ones with 'Reason_Disease' type are usually located in different sub-sentences, sometimes even in different sentences.

The above experimental results demonstrate that DiaKG is challenging for most current models and it is encouraged to employ more powerful models on this dataset.

5 Conclusion & Future Work

In this paper, we introduce DiaKG, a specified dataset dedicated to the diabetes disease. Through a carefully designed annotation process, we have obtained a high-quality diabetes dataset. The experiment results prove the practicability of DiaKG as well as the challenges for the most recent typical methods. We hope the release of this dataset can advance the construction of diabetes knowledge graphs and facilitate AI-based applications. We will further explore the potentials of this corpus and provide more challenging tasks like KBQA.

6 Acknowledgments

We want to express gratitude to the anonymous reviewers for their hard work and kind comments. We also thank Tianchi Platform to host DiaKG.

References

1. Li Y , Teng D , Shi X , et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study[J]. BMJ, 2020, 369. 1
2. Luo Z , Fabre G , Rodwin V G . Meeting the Challenge of Diabetes in China[J]. International Journal of Health Policy and Management, 2020, 9(2). 1
3. Holmes, David. Linong Ji: fighting to turn the tide against diabetes in China[J]. Lancet, 2014, 383(9933):1961-1961. 1
4. Nickel, M. et al. A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE 104.1(2015):11-33 1
5. Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. The American journal of sports medicine 2014;42(10):2371-6. 1
6. WangM, LiuM, LiuJ, et al. Safe medicine recommendation via medical knowledge graph embedding. arXiv preprint arXiv:1710.05980.2017. 1
7. Gann B. Giving patients choice and control: health informatics on the patient journey.[J]. Yearb Med Inform, 2012, 21(01):70-73. 1
8. X Li, Feng J , Meng Y , et al. A Unified MRC Framework for Named Entity Recognition[J]. 2019. 3.1

9. Peng Z , Wei S , Tian J , et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016. 3.2