AsdKB: A Chinese Knowledge Base for the Early Screening and Diagnosis of Autism Spectrum Disorder

Tianxing Wu[1,2], Xudong Cao[1], Yipeng Zhu[1], Feiyue Wu[1], Tianling Gong[1],
Yuxiang Wang[3], and Shenqi Jing[4,5]

[1] Southeast University, China
{tianxingwu, xudongcao, yipengzhu, wufeiyue, gtl2019}@seu.edu.cn
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its
Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3] Hangzhou Dianzi University, China
lsswyx@hdu.edu.cn
[4] The First Affiliated Hospital of Nanjing Medical University, China
[5] Nanjing Medical University, China
jingshenqi@jsph.org.cn

Abstract. To easily obtain the knowledge about autism spectrum disorder and help its early screening and diagnosis, we create AsdKB, a Chinese knowledge base on autism spectrum disorder. The knowledge base is built on top of various sources, including 1) the disease knowledge from SNOMED CT and ICD-10 clinical descriptions on mental and behavioural disorders, 2) the diagnostic knowledge from DSM-5 and different screening tools recommended by social organizations and medical institutes, and 3) the expert knowledge on professional physicians and hospitals from the Web. AsdKB contains both ontological and factual knowledge, and is accessible as Linked Data at **https://w3id.org/asdkb/**. The potential applications of AsdKB are question answering, auxiliary diagnosis, and expert recommendation, and we illustrate them with a prototype which can be accessed at **http://asdkb.org.cn/**.

Keywords: Autism Spectrum Disorder · Knowledge Base · Ontology.

1 Introduction

Autism spectrum disorder (ASD) is a kind of neurodevelopmental disability which begins before the age of 3 years and can last throughout a person's whole life. People with ASD have problems in social communication and interaction, and may have stereotypic or repetitive behaviors (or interests). According to the most recent statistics [17] published by the Centers for Disease Control and Prevention (CDC), about 1 in 36 children aged 8 years has been identified with ASD, and this proportion is quite high. However, there is no quantitative medical test to diagnose such a disorder, and professional physicians only use screening tools and look at the behaviors for some time to make a diagnosis. In
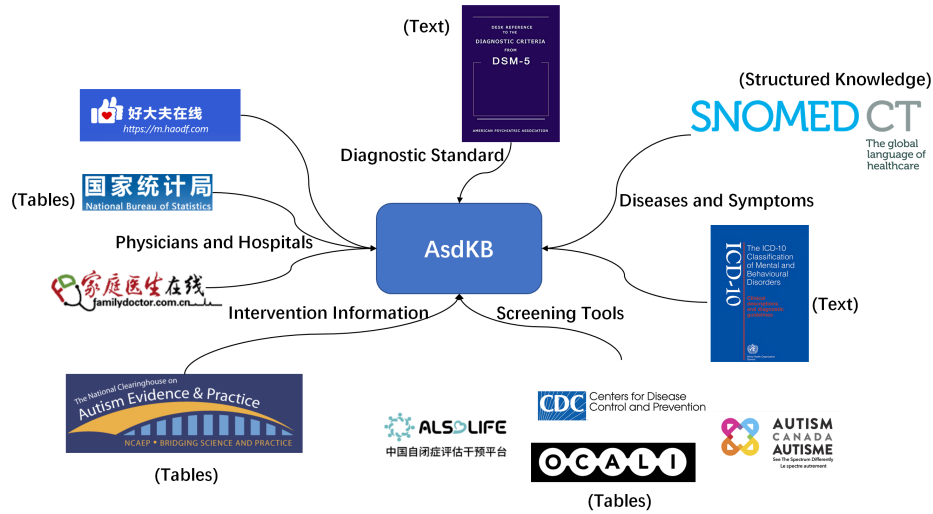
Fig. 1. The data sources for building AsdKB.

this way, many children cannot receive a final diagnosis until much older, which causes the children with ASD might not get the early help they need. In China, the situation on screening and diagnosing the children with ASD maybe much worse compared with western developed countries. The 2020 China rehabilitation report of children developmental disorder[6] points out that the ASD incidence in China is around 1% and the number of ASD children is more than three million, but the number of professional physicians who can diagnose ASD is only about 500, let alone the number of board certified behavior analysts. This does hinder the timely diagnosis on ASD, which inspires us to think about if we can apply artificial intelligence techniques to solve the early screening and diagnosis of ASD. The key problem is how to extract and integrate ASD relevant knowledge from heterogeneous sources to support upper-level intelligent applications.

To solve this problem, we build AsdKB, a Chinese knowledge base for the early screening and diagnosis of ASD, from various sources (see Figure 1), such as SNOMED CT [5] (a large collection of medical terms), ICD-10[7] (the 10th revision of the classification system of diseases published by WHO) clinical descriptions on mental and behavioural disorders [21], DSM-5 [1] (the 5th edition of diagnostic and statistical manual of mental disorders), the screening tools recommended by CDC and so on. Specifically, we first build an ontology covering important concepts about the screening and diagnosis of ASD from DSM-5, ICD-10 clinical descriptions on mental and behavioural disorders, SNOMED CT, CDC materials, and other Web sources. Using this ontology as the schema, we then extract and integrate factual knowledge on diseases, diagnosis, experts, and

---

[6] http://pkucarenjk.com/news-family/2303.html
[7] https://en.wikipedia.org/wiki/ICD-10

others. Besides, we use and develop Web crawler and natural language processing (NLP) tools for data extraction, keyword extraction, knowledge extraction, machine translation, and etc., over various formats of data, including text, tables, and structured knowledge. All classes, properties, and instances in AsdKB are identified by permanent dereferenceable URIs in w3id[8]. All data are available as RDF dump files on Zenodo[9], and the basic information of the AsdKB project can be accessed at Github[10]. All the resources are published under CC BY-SA 4.0. The main contributions of this paper are summarized as follows:

- We first build a Chinese knowledge base for the early screening and diagnosis of ASD, i.e., AsdKB, which contains both ontological and factual knowledge, and publish it following Linked Data best practices.
- We present a prototype system on question answering, auxiliary diagnosis, and expert recommendation with AsdKB, and discuss how to support the early screening and diagnosis of ASD with this system.

The rest of this paper is organized as follows. Section 2 introduces the process of ontology building. Section 3 describes the extraction of factual knowledge. Section 4 presents the potential applications of AsdKB. Section 5 outlines related work, and we conclude in the last section.

2   Ontology Building

This section introduces the process of building the AsdKB ontology as the schema which is used to guide extracting and integrating factual knowledge from various sources. We follow Ontology Development 101 [20] to build the ontology (Figure 2 shows a part of it) as follows.

Step 1: Determine the domain and scope of the ontology. AsdKB is expected to cover the ASD relevant knowledge on the early screening and diagnosis, so the ontology needs to cover important concepts in widely recognized materials about the screening and diagnosis of ASD. Here, we select relevant materials from CDC, DSM-5, ICD-10, SNOMED-CT, and other Web sources.

Step 2: Consider reusing existing ontologies. In this part, we reuse the standard RDF, RDFS, and OWL vocabularies, including `rdf:type` linking from instances to classes, `rdfs:label` recording the Chinese (or English) labels of classes and properties, `rdfs:comment` providing textual descriptions to clarify meanings of classes, `rdfs:subClassOf` describing the class hierarchy, equivalent classes are linked by `owl:equivalentClass` from the AsdKB ontology to other ontologies, and `rdfs:domain` and `rdfs:range` specifying the resources and values of a property are instances of one or more classes, respectively.
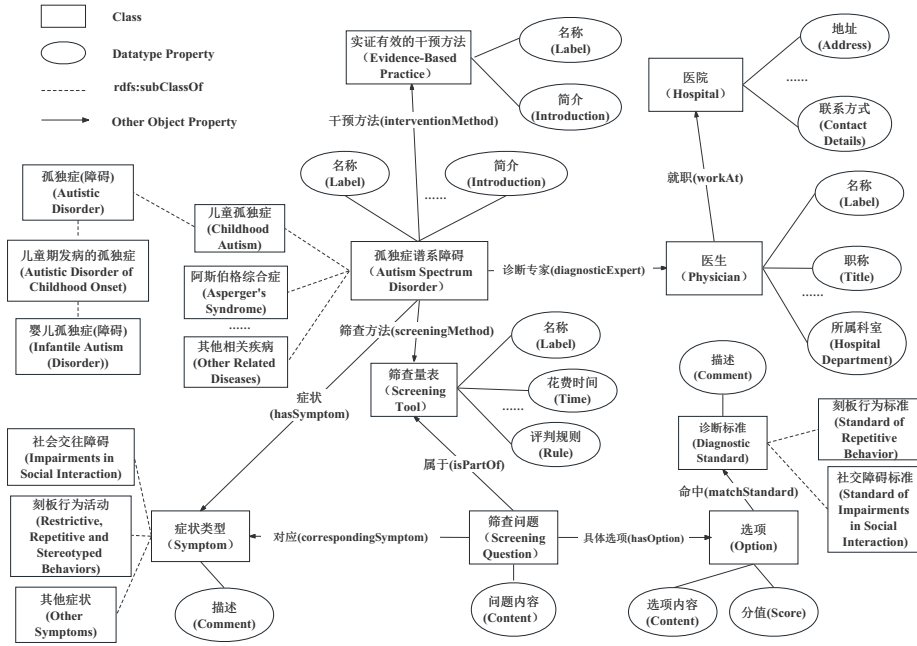
---

[8] https://w3id.org/asdkb/
[9] https://zenodo.org/record/8199698
[10] https://github.com/SilenceSnake/ASDKB

Fig. 2. A part of the AsdKB ontology.

Step 3: Enumerate important terms in the ontology. We read the ASD materials from CDC, DSM-5, ICD-10, SNOMED CT and other Web sources mentioned in the first step, to manually identify a list of important concept-level terms. For example, important symptom concepts in disease knowledge include "Impairments in Social Interaction" and " Restrictive, Repetitive and Stereotyped Behaviors". Important concepts in expert knowledge include "Physician" and "Hospital". Besides, "Screening Tool" and "Diagnostic Standard" are related to screening and diagnosis.

Step 4: Define the classes and the class hierarchy. Based on the previous identified important terms, we start to create disease classes (e.g., "Autism Spectrum Disorder" and "Asperger's Syndrome"), diagnosis classes (e.g., "Screening Tool" and "Screening Question"), expert classes (e.g., "Physician" and "Hospital"), and others. For the class hierarchy, we consider the hierarchies within disease classes, symptom classes, and diagnosis classes, respectively. For example, as shown in Figure 2, we have " Asperger's Syndrome `rdfs:subClassOf` Autism Spectrum Disorder" and "Standard of Social Interaction `rdfs:subClassOf` Diagnostic Standard". Specifically, we have created a class "Screening Question" in the diagnosis classes to facilitate the exploration of the association between instances of "Screening Question" and "Diagnostic Standard".

Step 5: Define the properties of classes. After selecting classes from the list of terms, we start to attach properties to classes using `rdfs:domain`. We distinguish datatype properties and object properties. For example, for the class "Physician", we have the object property `workAt` and datatype properties `Name`, `Title`, `Specialty`, `Hospital Department`, and etc.

Step 6: Define the facets of the properties. We specify the value type of each property by defining `rdfs:range`. The range of a datatype property is an XML Schema datatype. For example, the ranges of properties `Address` (attached to the class "Hospital") and `Score` (attached to the class "Option") are `xsd:string` and `xsd:float`, respectively. Besides, the range of an object property is a class. For example, the range of `hasSymptom` is the "Symptom".

Step 7: Create instances. We do not define instances in the ontology but only use it as the schema of AsdKB. The creation of instances belongs to factual knowledge extraction, and it will be described in Section 3.

Statistics about the AsdKB ontology. The built ontology is currently online: **https://w3id.org/asdkb/ontology/**. It contains 32 classes, 25 datatype properties, and 16 object properties. The maximum depth of a class in the class hierarchy is 4. Note that we apply google translate[11] to translate the English labels of all elements in the ontology into Chinese ones, and also perform careful manual proofreading and correction.

Mapping to other ontologies. To facilitate schema knowledge sharing across different ontologies, We map our AsdKB ontology to the Unified Medical Language System [2] (UMLS) and the Autism DSM-ADI-R (ADAR) ontology [19]. UMLS is the largest integrated biomedical ontology covering the vocabularies from around 200 sources including SNOMED CT, DSM-5, FMA [23], and etc. ADAR is built from Autism Diagnostic Interview-Revised [16] (ADI-R), which is a structured interview used for autism diagnosis. ADAR focuses on autism symptom classification, and it constructs many fine-grained symptom classes, e.g., "First walked unaided" and "Daily spontaneous and meaningful speech", which are taken as instances in AsdKB. This is why we do not directly re-use ADAR in AsdKB.

Since the disease classes in the AsdKB ontology are extracted from SNOMED CT, which is also a part of UMLS, such classes are naturally linked to UMLS (see the Example 1 in Figure 3). For the rest eighteen classes in AsdKB, we submit each of their labels to UMLS Metathesaurus Browser[12] and manually make comparisons between the returned classes and submitted ones to decide whether there exist `owl:equivalentClass` or `rdfs:subClassOf` relations (the Example 2 in Figure 3 gives a mapping result). Besides, we apply Agreement-MakerLight [6] to mapping the AsdKB ontology to ADAR, and the Example 3 in Figure 3 also shows a mapping result.

---

[11] https://translate.google.com/
[12] https://uts.nlm.nih.gov/uts/umls/home

Example 1:
<https://w3id.org/asdkb/ontology/class/autism_spectrum_disorder>   owl:equivalentClass
<https://uts.nlm.nih.gov/uts/umls/concept/C1510586> .

Example 2:
<https://w3id.org/asdkb/ontology/class/physician>   rdfs:subClassOf
<https://uts.nlm.nih.gov/uts/umls/concept/C0031831> .

Example 3:
<https://w3id.org/asdkb/ontology/class/restrictive_repetitive_and_stereotyped_behaviors>
owl:equivalentClass
<http://purl.org/autism-ontology/class/1.0/autism-rules.owl#Restricted_and_Repetitive_Behavior> .

Fig. 3. Examples of mapping the AsdKB ontology to UMLS and ADAR.

3   Factual Knowledge Extraction

This section presents the extraction of factual knowledge of ASD. Due to limited spaces, we do not explain every detail but focus on the main content.

3.1   Disease Knowledge

For disease knowledge, we need to extract the factual knowledge about disease and symptom instances according to the AsdKB ontology. Disease instances (e.g., "Atypical Rett syndrome") are derived from SNOMED CT, and they are actually the leaf nodes in the disease taxonomy in SNOMED CT. For each disease instance, we extract the values of the properties: `Label` (instance name), `SCTID` (the term ID in SNOMED CT), `ICD-10 code` (the corresponding ICD-10 category), and `Synonym` from SNOMED CT, respectively. We also manually design templates (i.e., regular expressions) to extract the values of properties `Introduction` (a brief description of the given disease instance), `Patient Groups` (e.g., "children" or "female children"), and `Pathogeny` (e.g., "genetic and environmental factors") from ICD-10 clinical descriptions on mental and behavioural disorders [21], respectively. Besides, for the values of properties `Label`, `Synonym`, `Introduction`, `Patient Groups`, and `Pathogeny`, we obtain the corresponding Chinese versions by Google Translate and manual proofreading. We collect 49 disease instances relevant to ASD in total, and their corresponding property information.

Symptom instances are also extracted from ICD-10 clinical descriptions on mental and behavioural disorders. We model the symptom instance extraction as the task of sequence labeling. We first take each paragraph as a document, and apply Term Frequency-Inverse Document Frequency [15] (TF-IDF) to identify keywords. Based on this, we then label a small amount of symptom instances in the corpus to train an extraction model. Here, we use BioBERT [14], a pre-trained biomedical language representation model for biomedical text mining, to encode each word as an embedding. Afterwards, we utilize BiLSTM [7] to capture textual context features for each word. Finally, we apply conditional

random fields [13] to finishing sequence labeling, which naturally classifies symptom instances to the pre-defined symptom classes, i.e., "Impairments in Social Interaction", " Restrictive, Repetitive and Stereotyped Behaviors", and "Other Symptoms". High-quality results of sequence labeling obtained by the trained model will be added to the labeled data to train a new model. We repeat this process until the maximum number of iterations is reached. Google Translate is also used to get the Chinese description of each symptom instance. Figure 4 shows the triples of the symptom `<https://w3id.org/asdkb/instance/symptom64>`. Finally, We collect 65 symptom instances in total.

```
<https://w3id.org/asdkb/instance/symptom64>    rdfs:comment
"a resistance to changes in routine or in details of the personal environment (such as the movement
of ornaments or furniture in the family home)."@en .

<https://w3id.org/asdkb/instance/symptom64 >    rdfs:comment
"对日常生活或个人环境细节变化的抵制（如家庭中装饰品或家具的移动）"@zh .

<https://w3id.org/asdkb/instance/symptom64 >    rdf:type
<https://w3id.org/asdkb/ontology/class/restrictive_repetitive_and_stereotyped_behaviors> .
```

Fig. 4. An example of the triples containing a symptom instance.

3.2   Diagnostic Knowledge

For diagnostic knowledge, we extract the factual knowledge on the instances of diagnostic standards, screening tools, screening questions, and the corresponding options. Instances of diagnostic standards are acquired from the Chinese edition[13] of DSM-5 [1], so we only have Chinese descriptions for the instances of diagnostic standards. We follow a similar process used for extracting symptom instances, and only replace the pre-trained model BioBERT with a more general model BERT [11] because BioBERT does not support Chinese but BERT does. Different from symptom instances, instances of diagnostic standards do not refer to specific behaviors or activities of the people with ASD, they actually present textual summarizations for specific classes of diagnostic standards (i.e., "Standard of Repetitive Behavior" and "Standard of Impairments in Social Interaction"). For example, an instance of diagnostic standards in AsdKB is expressed as " 眼神接触、手势、面部表情、身体定位或言语语调等方面的缺乏、减少或不合规的使用 (abnormalities in eye contact and body language or deficits in understanding and use of gestures)", which corresponds to multiple symptom instances, e.g., " 不会进行对视 (without eye contact)" and "很少微笑 (rarely smile)". We collect 43 instances of diagnostic standards in total.

Regarding instances of screening tools, screening questions, and options, we extract them and the corresponding property values from the websites of social

---

[13] https://www.worldcat.org/zh-cn/title/1052824369

organizations and medical institutes, including CDC[14], ALSOLIFE[15] (China ASD Evaluation and Intervention Platform), Autism Canada[16], and OCALI[17] (The Ohio Center for Autism and Low Incidence). Instances of screening tools in AsdKB are actually screening scales, which have the properties `Introduction` (basic information and instructions), `Author`, `User` (the one filling in the scale, e.g., parents or teacher), `Age` (applicable ages of screening targets), `Time` (the time it takes to fill in the scale), `Rule` (screening principles and details), and `Screening Boundary` (the score of screening boundary after finishing the scale). After careful selection, we extract twenty instances of screening tools, containing fifteen English screening scales and five Chinese ones, such as ABC [12], CARS2 [24], and M-CHAT [27]. Google Translate is used here to translate English scales into Chinese ones, and manual proofreading is also conducted.

For instances of screening questions and options, they can be directly obtained from screening scales through table extraction. Besides keeping their textual content as the property, we also establish `correspondingSymptom` relationships between instances of screening questions and symptom instances, and `matchStandard` relationships between option instances and instances of diagnostic standards. These two kinds of relationships (i.e., object properties) benefit to the interpretability of screening results. For example, as shown in Figure 5, AsdKB can tell users that the current question investigates what specific symptoms are and whether the current option matches some diagnostic standard or not, in order to help users better understand screening questions and provide explanations to screening results. To identify `correspondingSymptom` relationships, we first use FNLP [22] to perform Chinese word segmentation on the instances of screening questions and symptom instances. After removing stopwords, we then compare string similarities between two word sequences to decide whether the `correspondingSymptom` relationship exists. The method of extracting `matchStandard` relationships is similar to that of `correspondingSymptom` relationships, and the only difference is to additionally consider the property `Score` of each option. If an option has the highest score or lowest score, it means the result of the current screening question is abnormal or normal (it ups to the design of screening scales), and abnormal results could help identify `matchStandard` relationships.

3.3   Expert Knowledge

For expert knowledge, we extract factual knowledge of the instances of professional physicians, and the hospitals they work at, from the Web. We select two famous Chinese healthcare websites The Good Doctor[18] and Family Doctor[19]
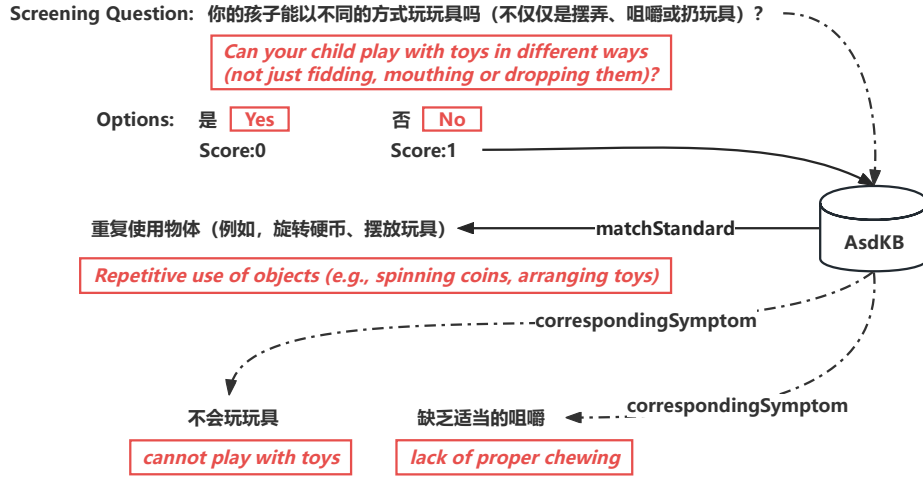

Fig. 5. An example to show the benefits of `correspondingSymptom` and `matchStandard` relationships.

as the data sources to extraction, so the string values of some datatype properties are only presented in Chinese. We first submit the following keywords "孤独症, 自闭症, 孤独症谱系障碍 (ASD)", "广泛性发育障碍 (pervasive developmental disorders)", "儿童孤独症 (childhood autism)", and "阿斯伯格综合症 (Asperger's syndrome)", to the search engines of the selected websites, which locates the Web pages of professional physicians on ASD. Faced with the structures like infobox tables in Wikipedia, we then extract physician instances and the values of properties `Name`, `Title` (e.g., "主任医师 (chief physician)" and "主治医师 (attending physician)"), `Specialty` (e.g., "各类儿童精神障碍 (various types of mental disorders in childhood)"), `Hospital Department` (e.g., "儿童保健科 (child healthcare department)" and "精神科 (psychiatry department)"), and `workAt` (i.e., hospital instances). We collect 499 physician instances in total.

According to the values of the property `workAt`, we locate the Web pages of hospital instances. Similar to the extraction on physician instances and the corresponding property information, we extract the hospital instances and the values of properties `Name`, `Address`, `Contact Details`, and `Hospital Level` (e.g., "三甲医院 (Grade-A tertiary hospital)"). We collect 270 hospital instances in total.

Since physician and hospital instances are extracted from different sources, we perform instance matching using heuristics. Given two hospital instances, if their values for at least one of the properties `Address` and `Contact Details` are the same, they are treated as equivalent. Given two physician instances, if their values for the property `workAt` are equivalent, and the values for the properties `Name` and `Title` are the same respectively, these two instances are determined as equivalent. Equivalent instances are fused as one instance in AsdKB.

3.4   Other Knowledge

In this part, we extract factual knowledge on the instances of intervention methods, and the information of China administrative divisions. Instances of intervention methods are obtained from The National Clearinghouse on Autism Evidence and Practice[20] (NCAEP), and such instances are all evidence-based practices, including "Discrete Trial Training", "Social Skills Training", "Peer-Based Instruction and Intervention", and etc. For each instance of intervention methods, we extract the values of properties `Label` (instance name) and `Introduction` (a brief description on the instance information). English string values are translated to Chinese by Google Translate, and we also conduct careful proofreading.

With expert knowledge introduced in Section 3.3, a potential application is to seek expertise help from physicians to diagnosis. In order to find professional physicians in the target districts, cities, and provinces, we extract instances of China administrative divisions from National Bureau of Statistics[21]. The extracted instances are specific districts, cities, and provinces, and we also build `locateAt` relationships among them. To link each hospital to the corresponding administrative divisions, we first use Amap (a leading provider of digital map in China) API[22] to get the latitude and longitude of each hospital by inputting the value of property `address`. With the information of latitudes and longitudes, Amap API can return the corresponding districts, cities, and provinces of hospitals. Besides, we record the `Population` of each instance of China administrative divisions, which could help regional analysis on ASD.

3.5   Quality of AsdKB

AsdKB contains 6,166 entities (including conceptual entities, i.e., classes, and individual entities, i.e., instances) and 69,290 triples in total. All class URIs in the namespace **http://w3id.org/asdkb/ontology/class/** and instance URIs in the namespace **http://w3id.org/asdkb/instance/** are dereferenceable. To evaluate the quality of AsdKB, we design two evaluation methods: accuracy evaluation, and task evaluation.

Accuracy Evaluation. There is no ground truth available, and it is impossible to evaluate all triples manually. Therefore, we apply a random evaluation strategy. We first randomly select 100 entities distributed across classes and instances, and obtain 732 triples. These samples can reflect the distribution of triples in the entire knowledge base. We then conduct manual labeling to evaluate the accuracy of the samples. The accuracy of the entire AsdKB is estimated by evaluating the accuracy of the samples.

Five graduate students participate in the labeling process. We provide three choices, which are *correct*, *incorrect*, and *unknown* to label each sample. After each student label all the samples, we calculate the average accuracy. Finally,

---

[20] https://ncaep.fpg.unc.edu/
[21] http://www.stats.gov.cn/
[22] https://github.com/amapapi

similar to YAGO [10], Zhishi.me [28], and Linked Open Schema [29], we use the Wilson interval [3] when $\alpha = 5\%$ to extend our findings on the subset to the entire knowledge base. The Wilson interval is a binomial proportion confidence interval calculated from the results of a series of Bernoulli trials, and α is the significance level. For the randomly selected 732 triples, the average *correct* votes is 712, so the accuracy is $97.02\% \pm 1.21\%$, and it demonstrates the high quality of AsdKB.

Task Evaluation. Besides the accuracy of the triples in AsdKB, we try to evaluate the effectiveness of AsdKB in answering real-world ASD relevant questions. Thus, we collect 100 frequently asked questions (e.g., "孤独症都有哪些临床表现？ (What are the clinical symptoms of autism?)" and "哪些干预方法是有效的？ (Which interventions are effective?)") on ASD from Chinese healthcare websites The Good Doctor and Family Doctor (introduced in Section 3.3), which are also the data sources of the expert knowledge in AsdKB. We store AsdKB in a graph database Neo4j [26], and also invite five graduate students to manually write Cypher (Neo4's graph query language) queries for the collected questions so as to check whether the returned query results can answer the questions. According to the above evaluation, AsdKB can answer 81 questions, i.e., the coverage reaches to 81%, which reflects the practicality of AsdKB.

4   Application of AsdKB

To illustrate the potential application of AsdKB, this section describes the implementation of a prototype system[23] for the early screening and diagnosis of ASD based on AsdKB. This system has three main applications, including question answering, auxiliary diagnosis, and expert recommendation. Users of this system are parents, teachers, and caregivers.

4.1   Question Answering

We implement a natural language question answering (QA) system based on AsdKB, and expect that the QA system can answer various common-sense and factual questions on ASD. As mentioned in Section 3.5, AsdKB is stored in Neo4j, so we aim to translate each natural language question to a Cypher query, in order to query the graph database to return the answer. We use two strategies to design the QA system. The first one is to manually write common ASD relevant natural language query patterns (i.e., regular expressions) according to AsdKB ontology and the corresponding Cypher query templates. If a user query matches one of our patterns, then we construct and execute the corresponding Cypher query based on the pre-defined Cypher query template to get the answer. If the user query does not match our patterns, we use the second strategy, which applies the idea of the method for translating natural language questions to formal queries with semantic query graph modeling [31] to generating the Cypher query.
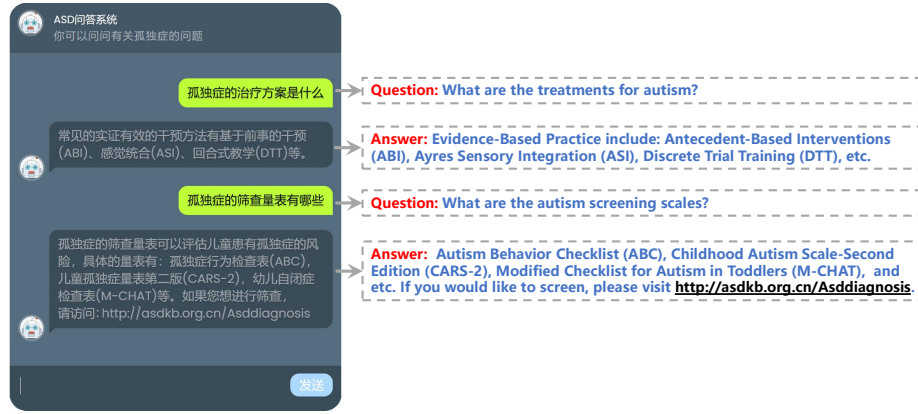
---

[23] http://asdkb.org.cn/

Fig. 6. The interface of the QA system.

Figure 6 shows the interface of our QA system. We also detect the intention of each question to check whether the user would like to further fill in screening scales. If so, the system will directly give the link of auxiliary diagnosis to help choose screening scales (see Figure 6). The intention identification is modeled as a task of binary classification, where we use BERT to encode questions in the labeled data, and then train a SVM [9] classifier to predict whether users are willing to conduct screening or not.

4.2   Auxiliary Diagnosis

We have developed an auxiliary diagnosis system based on AsdKB. This system provides users with screening scales to assess the risk of being ASD. As long as the screening result of a screening scale shows a risk, the system will prompt the user to seek professional medical evaluation and recommend experts using our expert recommendation system (will be introduced in Section 4.3).

As shown in Figure 7(a), before filling the screening scales, users can select appropriate screening conditions based on their situations, such as the child's age and existing symptoms, and the system will return the corresponding screening scales with a brief introduction (see Figure 7(b)). Figure 7(c) shows the questions and options when filling in the ABC screening scale. After completing a screening scale, the system will give the screening result (i.e., risky or not) based on the total score of all options and the screening boundary.

When users are filling in screening scales, they can check what specific symptoms the current question investigates to better understand the question, so as to help make a choice more precisely. Besides, after completing screening scales, this system can also analyze which option matches some diagnostic standard, to provide explanations of the screening results. More details have already been introduced in Section 3.2 and Figure 5.

Fig. 7. An illustration of the auxiliary diagnosis system.

4.3 Expert Recommendation

If our auxiliary diagnosis system reports the risk of being ASD, users may have requirements to find experts on diagnosing ASD in the target administrative divisions. Thus, we design an expert recommendation system with facet search on AsdKB. Users can choose the target province, city and district by selecting a checkbox or directly clicking their locations on the map (see Figure 8). The recommendation result is a list of professional physicians with their names, titles, hospital departments, hospitals, hospital addresses, and specialties.

The recommendation has two steps: candidate physician generation and candidate physician ranking. In candidate physician generation, we use the location information of hospitals in AsdKB to match user selected administrative divisions, and the physicians in AsdKB working at such hospitals are candidates. Note that if no candidate physician returns, we will consider more hospitals in surrounding administrative divisions by distance calculation with latitudes and longitudes. In the candidate physician ranking, three aspects are taken into consideration. Firstly, the higher the title, the higher the ranking. Secondly, the higher the hospital level, the higher the ranking. Finally, the higher the number of thumbs up minus the number of thumbs down (Figure 8 gives an example), the higher the ranking.
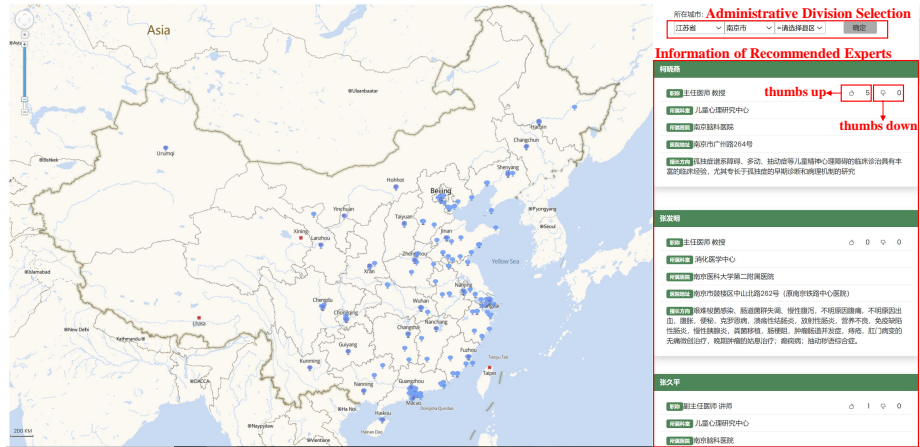
Fig. 8. An illustration of the expert recommendation system.

5   Related Work

Tu et al. [25] first proposed an autism ontology with domain terms and relationships relevant to autism phenotypes. The main target is to enable user queries and inferences about such phenotypes using data in the NDAR repository, but it does not include DSM criteria, so it does not support diagnosis of ASD. McCray et al. [18] also developed an ASD-phenotype ontology assessing and comparing different ASD diagnostic instruments, but it also does not include DSM-IV or DSM-5 criteria phenotypes. ADAR [19] extends an ontology proposed by Tu et al [25]. with additional SWRL rules to infer phenotypes from ADI-R [16] items, and it covers various symptoms and features of DSM IV and DSM-5 diagnostic criteria, such as difficulties with social interaction, language and communication issues, and stereotyped and repetitive behaviors. However, many fine-grained classes are actually instances in the generic sense.

The most recent work is AutismOnt [8], an ontology for autism diagnosis and treatment, which covers various covers autism research directions. AutismOnt includes the classes: Diagnosis, Risk Factors, Treatments, Strength and Weakness, Services, Lifespan Issues, Profile, and Family Relationships. However, although the authors claim that AutismOnt is available in the NCBO BioPortal, it cannot be found in the repository.

Some large-scale medical knowledge bases also contain ASD knowledge. For example, SNOMED CT [5] contains a large-scale number of medical terms, and the disease classes in AsdKB also comes from SNOMED CT, but it does not cover other kinds of knowledge, such as diagnostic knowledge and expert knowledge. Yuan et al. [30] proposed a method for constructing knowledge graphs with minimal supervision based on unstructured biomedical domain-specific contexts.

They collected 24,687 abstracts of articles related to ASD from PubMed[24], and constructed a knowledge graph on ASD. However, they did not design the ontology and the knowledge graph is not publicly available. CMeKG [4] is a Chinese medical knowledge graph developed using natural language processing and text mining techniques from a large amount of medical text data. CMeKG mistakenly uses drugs as the treatment for ASD, but drugs are only used to alleviate the complications of ASD in fact.

Compared with all existing works, AsdKB is the first publicly available Chinese knowledge base on ASD, and it contains both ontological and factual knowledge about diseases, diagnosis, experts, and others. AsdKB has been applied in developing applications of the early screening and diagnosis of ASD.

6   Conclusions and Future Work

We develop and publish a Chinese knowledge base on ASD called AsdKB by extracting and integrating knowledge from various data sources with different formats. To the best of our knowledge, AsdKB is the most comprehensive ASD knowledge base on the Web, and it supports the different applications on the early screening and diagnosis of ASD, such as question answering, auxiliary diagnosis, and expert recommendation. However, there are still some limitations to our work that we plan to address in the future.

Quality of AsdKB. During our preliminary evaluations of AsdKB, we discovered that the entities contained within the knowledge base are of high quality. However, errors do exist during the automatic extraction process. These errors stem from a variety of factors such as the quality of the original data sources, differences in data formats, and our integration methods. To address this issue, we plan to introduce crowd-sourcing techniques to fix the existing errors in AsdKB and study automatic error detection methods to ensure the accuracy of knowledge in the process of knowledge update.

Applications of AsdKB. We have explored various applications for AsdKB, including QA, auxiliary diagnosis, and expert recommendation. The integrated prototype system has demonstrated the potential for AsdKB to play a critical role in early ASD screening and diagnosis. To further improve the accuracy of QA and auxiliary diagnosis, we will incorporate data-driven machine learning models on more user log data in our prototype system. In addition to this, we plan to analyze electronic medical records if possible using AsdKB to assist physicians in ASD diagnosis. By analyzing medical histories, symptoms, and other relevant information using AsdKB, physicians can make more accurate diagnosis and give appropriate and personalised treatment suggestions to the people with ASD.

---

[24] https://pubmed.ncbi.nlm.nih.gov/

References

1. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders: DSM-5, vol. 5 (2013)
2. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. Brown, L.D., Cai, T.T., DasGupta, A.: Interval Estimation for a Binomial Proportion. Statistical Science **16**(2), 101–133 (2001)
4. Byambasuren, O., Yang, Y., Sui, Z., Dai, D., Chang, B., Li, S., Zan, H.: Preliminary Study on the Construction of Chinese Medical Knowledge Graph. Journal of Chinese Information Processing **33**(10), 1–9 (2019)
5. Donnelly, K.: SNOMED-CT: The Advanced Terminology and Coding System for eHealth. Studies in health technology and informatics **121**, 279 (2006)
6. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The AgreementMakerLight Ontology Matching System. In: Proc. of OTM. pp. 527–541 (2013)
7. Graves, A., Schmidhuber, J.: Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. Neural Networks **18**(5-6), 602–610 (2005)
8. Hassan, M.M., Mokhtar, H.M.: AutismOnt: An Ontology-Driven Decision Support For Autism Diagnosis and Treatment. Egyptian informatics journal **23**(1), 95–103 (2022)
9. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their Applications **13**(4), 18–28 (1998)
10. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial intelligence **194**, 28–61 (2013)
11. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. of NAACL. pp. 4171–4186 (2019)
12. Krug, D.A., Arick, J., Almond, P.: Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. Child Psychology & Psychiatry & Allied Disciplines (1980)
13. Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. of ICML. pp. 282–289 (2001)
14. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
15. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Data Sets. Cambridge university press (2020)
16. Lord, C., Rutter, M., Le Couteur, A.: Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of autism and developmental disorders **24**(5), 659–685 (1994)

17. Maenner, M.J., Warren, Z., Williams, A.R., Amoakohene, E., Bakian, A.V., Bilder, D.A., Durkin, M.S., Fitzgerald, R.T., Furnier, S.M., Hughes, M.M., et al.: Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. MMWR Surveillance Summaries **72**(2), 1 (2023)
18. McCray, A.T., Trevvett, P., Frost, H.R.: Modeling the Autism Spectrum Disorder Phenotype. Neuroinformatics **12**, 291–305 (2014)
19. Mugzach, O., Peleg, M., Bagley, S.C., Guter, S.J., Cook, E.H., Altman, R.B.: An ontology for Autism Spectrum Disorder (ASD) to infer ASD phenotypes from Autism Diagnostic Interview-Revised data. Journal of biomedical informatics **56**, 333–347 (2015)
20. Noy, N.F., McGuinness, D.L.: Ontology development 101: A Guide to Creating Your First Ontology. Technical report, Standford University (2001)
21. Organization, W.H., et al.: The ICD-10 Classification of Mental and Behavioural Disorders: clinical descriptions and diagnostic guidelines. World Health Organization (1992)
22. Qiu, X., Zhang, Q., Huang, X.J.: FudanNLP: A Toolkit for Chinese Natural Language Processing. In: Proc. of ACL. pp. 49–54 (2013)
23. Rosse, C., Mejino Jr, J.L.: The Foundational Model of Anatomy Ontology. In: Anatomy Ontologies for Bioinformatics: Principles and Practice, pp. 59–117 (2008)
24. Schopler, E., Van Bourgondien, M., Wellman, G., Love, S.: Childhood Autism Rating Scale–Second Edition (CARS-2). Los Angeles, CA: Western Psychological Services (2010)
25. Tu, S.W., Tennakoon, L., O'Connor, M., Shankar, R., Das, A.: Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: The Case of Autism. In: AMIA Annual Symposium Proceedings. vol. 2008, p. 727 (2008)
26. Webber, J.: A programmatic introduction to Neo4j. In: Proc. of SPLASH. pp. 217–218 (2012)
27. Wright, K., Poulin-Dubois, D.: Modified Checklist for Autism in Toddlers (M-CHAT): Validation and Correlates in Infancy. Comprehensive Guide to Autism **80**, 2813–2833 (2014)
28. Wu, T., Wang, H., Li, C., Qi, G., Niu, X., Wang, M., Li, L., Shi, C.: Knowledge graph construction from multiple online encyclopedias. World Wide Web **23**, 2671–2698 (2020)
29. Wu, T., Wang, H., Qi, G., Zhu, J., Ruan, T.: On building and publishing Linked Open Schema from social web sites. Journal of Web Semantics **51**, 39–50 (2018)
30. Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., Luo, J.: Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowledge and Information Systems **62**, 317–336 (2020)
31. Zou, L., Huang, R., Wang, H., Yu, J.X., He, W., Zhao, D.: Natural Language Question Answering over RDF: A Graph Data Driven Approach. In: Proc. of SIGMOD. pp. 313–324 (2014)