**This document includes all examples in Chinese with English translations and explanations in our paper. We have marked the corresponding sections and positions.**

* **Background**

***page 2 line 35***

For example, they may ask “ACEI (angiotension converting enzyme inhibitors)类药品使用患者的数目” (“The number of patients who use ACEI(angiotension converting enzyme inhibitors) drugs”.

* **Overview of the method**

***page 4 line 16***

For example, the original question is “患有高血糖没有患有高血压的病人吃了降血糖类药后葡萄糖化验结果偏低的女患者数量?” (“How many female patients with hyperglycaemia and no hypertension who have low levels of glucose tests after they took hypoglycemic drugs?”). The segmentation result of “女患者” (“female patients”) can be “女/患者” (“female / patients”) or “女患者” (“female patients”). “患有, 高血糖, 没有, 患有, 高血压, 的, 病 人, 吃, 了, 降血糖类药, 后, 葡萄糖, 化验结果, 偏低, 的, 女, 患者, 数量, ?” or “患有, 高血糖, 没有, 患有, 高血压, 的, 病人, 吃, 了, 降血糖类药, 后, 葡萄糖, 化验 结果, 偏低, 的, 女患者, 数量, ?” We pass all the possible results to later phase for disambiguation.

* **The representation of clinical data and the form of supported questions**

***page 7 line 31***

The questions “得了糖尿病后得心衰的患者” (“The patients who get diabetes and later suffer from heart failure”) can be regarded as etiological analysis. The drug usages are most frequently asked, such as “患有冠心病患者吃哪些中药?” (“What traditional Chinese medicine do the patients with coronary heart disease take? ”). At present, our system supports the following statistical types: lists, counts, aggregation (sum/averages), distributions, and ratios. The questions contain statistic vocabularies, such as “有哪些” (which), “比例” (ratio) and “分布” (distribution). Our system also supports questions with logical operations (and/or/not), such as “患有高血糖没有患有高血压的病人吃了降血糖类药后葡萄糖化验结果?” (“What is the glucose test result for patients with hyperglycaemia and no hypertension who have taken hypoglycemic drugs?” )

***page 8 line 5***

For the example, there are eight nodes and nine edges in “患有高血糖没有患有高血压的病人吃了降血糖类药后葡萄糖化验结果偏低的女性患者数量?” (“How many female patients with hyperglycemia and no hypertension have low levels of glucose tests after they took hypoglycemic drugs?” While the number of nodes and relations among clinical events give hints on the complexity of questions, we do not limit the length of the trees.

* **Word segmentation**

***page 8 line 18***

For example, the tool segments a traditional Chinese medicine “复 方甘草口服溶液” (compound glycyrrhiza oral solution) into four parts: “复方,” “甘草,” “口服,” and “溶液.”

***page 9 line 1***

To deal with different kinds of time-related expressions, such as “2015/6/10,” “2015年6月10日,” and “2015-6-10,” we define a series of rules to normalize time in a question into a format such as 20150610.

***page 9 line 8***

For example, “患有高血糖的女患者平均年龄是多少?” (What is the average age of female patients suffering from hyperglycemia?) The accurate mode of the segmentation result is “患有, 高血糖, 的, 女患者, 平均, 年龄, 是, 多少, ?”. However, “女患者” (female patient) is not an atomic concept that can be matched to the schema of the patient graph. In fact, the token contains two levels of information: “女” (female) is the value of the gender, and “患者” (patient) is the class patient. The correct segmentation result we need is “女/患者” (“female/patient”).

* **Concept linking**

***page 9 line 25***

The challenges arise from linking unknown tokens to classes, instances, relations, and properties. For example, the test item “血葡萄糖” (“blood glucose”) is often shortened as “葡萄糖” (“glucose”). Therefore, we use similarity matching to link the token to its possible concepts.

***page 10 line 6***

For example, on the token “葡萄糖” (glucose), we get the two top candidate words, “血葡萄糖” (blood glucose) and “葡萄糖口服溶液” (glucose oral liquid). While both the words belong to the instance set, they are of different types.

***page 10 line 16***

For example, the attribute name “年龄” (“age” in the question “60岁以上的高血糖女患者有多少?” (“How many female patients over 60 years old have hyperglycaemia?”) is omitted; only the attribute value “60岁” (60 years old) appears. Therefore, we use the schema information to supplement the missing attributes since they can cause grammar parsing in the later step to fail. For the previous example, the unit “岁” (“years old”) corresponds to the attribute “年龄” (“age” in our schema graph, and “女” (“female”) is an enumeration value of attribute “性别” (“sex”).

* **Context-free grammar**

***page 11 line 17***

Relation-condition chunks can consist of relationships and instances, such as “患有高血糖” (suffering from hyperglycemia), in which “患有” (“suffering from”) indicates a relationship and “高血糖” (“hyperglycemia”) represents an instance. There is a hidden concept “患者” (“patient”) in the relationship. However, users usually put it in the end of the condition when multiple condition exists. It can also be negative relation-condition chunks, such as “没有患有高血糖” (not suffering from hypertension).

***page 12 line 5***

For example, “葡萄糖化验结果偏低” (glucose test results are low), in which the “葡萄糖”(“glucose”) is an instance, “化验结果” (“test results”) belongs to attributes, and “偏低” (“low”) belongs to enumeration values. The subject can also be omitted, for example, “住院天数超过10天” (the day of stay in hospital exceeds 10 days), where “住院天数” (“the day of stay in hospital”) is an attribute and “超过10天” (“exceeds 10 days”) indicates numeric values of attributes in a range form.

***page 12 line 13***

The query-item chunks represent the statistical operations on the records that satisfy the combination condition. For example, the question about etiology “患有高血压，糖尿病，冠心病的病人发生心衰的比例?” (“What is the ratio of heart failure in patients with hypertension, diabetes, and coronary heart disease?”) uses the ratio operation. The question about effectiveness “患有高血糖没有患有高血压的病人吃了降血糖类药后葡萄糖化验结果偏低的女患者数量?” (“How many female patients with hyperglycemia and no hypertension have low levels of glucose tests after they took hypoglycemic drugs?”) contains several logic and time operators to filter the patient’s records and uses the count operation.

* **Using Stanford parser as complement**

***page 12 lin 25***

For example, “患有高血糖和高血压的病人” (“patients who suffer from hyperglycemia and hypertension”), our grammar cannot capture this relationship because of the remote dependency of “suffer from” and “hypertension.”

* **Semantic representation with disambiguation**

***page 13 line 6***

For instance, in “吃，感冒灵” (eat Coldrine)，“吃” (eat) is binded with the predicate “服用” (take) since the schema graph contains a synonym relation < 吃, sameAs，服用 >. In the concept linking step, “感冒灵” can be linked to “感冒” (cold) and ” 感冒灵胶囊” (Coldrine capsule). Based on the patient-schema graph, < 病人(patient), 患有(suffer from), 感冒 >、< 病人(patient), 服用 (take), 感冒灵胶囊 >, the latter is chosen since the predicate is ” 服用” (take). The property-conditional chunks are similar to relation-condition chunks. Our query combines statistic operators with classes and property values. With the help of query-item chunks, we can find out the object that the statistic operator works on.

* **Cypher translation**

***page 13 line 33***

For example, the patient node directly connects with the disease node, the test node, and the drug node; and the structure corresponds to the match clause in Figure 4. The gender, disease, and drug constraints of the where clause in Figure 4 correspond to the attribute constraints of class-node patients (患者), whereas the time constraint between “吃了” (“take”) and “做化验” (“test”) maps from TimeOperator in the graph. The return clause ultimately returns the number of patients who satisfy the subgraph.

* **Experimental evaluation on the coverage and accuracy of the system**

***page 14 line 14***

We search literature on Wanfang Data, a literature-resource library with major research journals in Chinese, using different combinations of three groups of key words: 电子病历 (electronic medical records)/电子健康档案 (electronic health records), 统计 (statistical)/数据分析 (data analysis), and 流调 (epidemiological investigation)/临床研究 (clinical research).

***page 16 line 17***

The remaining 46 questions whose results are not returned can be divided into four categories:

1.There is one question with a statistical vocabulary “年增长率” (annual growth rate) that is not supported by the system. We can solve this problem by adding this word to the statistical vocabulary and setting up its related function.

2.There are 17 questions whose semantics cannot be interpreted directly, for example, “1991年与2014年居民高血压患病率对比” (What is the comparison of the prevalence of hypertension between 1991 and 2014?). Currently, the system’s grammar cannot parse and interpret the term “对比” (comparison) correctly. However, we can extend the vocabulary and parse them into two tree-like structures later.

3.Two questions contain operations not supported in Neo4j, for example “年住院患者前10位疾病性别构成” (“What is the sex ratio of the top 10 diseases among inpatients in 2012?” We need to sort the group by operation for each disease first, but it is not supported in Neo4j.

***page 17 line 12***

1. Questions, which collected from doctor, contain vocabularies that do not exist in our knowledge base. For example, “患有败血症、心衰和冠心病的病人总数” (“ How many patients are suffering from sepsis, heart failure, and coronary heart disease?”, “sepsis” is not in our knowledge base. Our system automatically ignores the entities, so the result is the same as “患有心衰和冠心病的病人总数” (“ How many patients are suffering from heart failure and coronary heart disease?” To solve the problem, we plan to add a named entity recognition step. If the entity cannot link to any entity in the knowledge base, we will prompt the user.

2. Entities in the question are linked to wrong entities in the knowledge base. Entities in the question may not exist in the knowledge base, so they are linked to the wrong entities, for example, “得了冠心病和风湿性心脏病的患者有哪些” (“Who are the patients with coronary heart disease and rheumatic heart disease?” The system will link rheumatic heart disease to heart disease. In this case, a larger knowledge base is required.

3. Our system can only recognize the time relation between two entities where temporal preposition exists. In “患有心衰后又得了冠心病的高血压患者” (“High blood pressure patients with coronary heart disease after heart failure”, our system can only know that heart failure occurs before coronary heart disease, but the hidden semantics that hypertension happens earlier than the previous two cannot be captured. For this case, we intend to define a set of heuristic rules to recognize such domain-specific implications.