# DLUT Question Answering System at NTCIR-8 CCLQA

Wei Wang
School of Computer Science & Technology, Dalian University of Technology
1-12-3, No.8, XiangLong Street GanJingZi District, Dalian, China
0086-411-87185161
wangwei.dl@hotmail.com

Degen Huang，Lishuang Li
School of Computer Science & Technology, Dalian University of Technology
No.2 Linggong Road, Ganjingzi District, Dalian, China
0086-411-84708140, 84706002-3930
huangdg@dlut.edu.cn, computer@dlut.edu.cn

Xiuying Li
School of Foreign Languange, Dalian University of Technology
No.2 Linggong Road, Ganjingzi District, Dalian, China
xyli1@163.com

## ABSTRACT

*This paper describes the Chinese question answering system DLUT for the CS-CS Subtask evaluation of NTCIR-8 CCLQA. The system utilizes Situation Unit (SU), combined syntactic and semantic information, as a basic processed unit to get candidate answers, wherein the SUs in the question sentences are matched with those in the texts of corpus. In this evaluation, answers are presented in the form of whole sentences instead of their simplified versions. The average F3 Score reaches 0.1954, and the average Recall is ≥ 0.75 for the six question types, BIOGRAPHY, DATE, DEFINITION, LOCATION, ORGNIZATION and PERSON. As the current system only employs the partial information in SUs, the evaluation result only indicates that the SU-based question answering system can get promising Recall.*

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]:Retrieval models and Search process

H.3.4 [**Systems and Software**]:Question-answering (fact retrieval) systems

## General Terms

Algorithms, Theory

## Keywords

Information Retrieval, Question Answering System, Situation Unit (SU)

## 1. INTRODUCTION

The question answering system is a further development of the conventional information retrieval techniques. Most of the current question answering systems are built on syntactic and semantic information. Boris Katz, Gary Borchardt, et al. integrated syntactic and semantic interpretation approach into their system START[1]; Hyeon-Jin Kim, Ji-Hyun Wang, et al. proposed to construct Answer Index Unit (AIU) and then computed answer candidate weight by taking into account syntactic information [2]; Powerset conducted syntactic analysis for every sentence on the webpage and extracted entities and semantic information to generate index [3]; PARC builds the Bridge system to realize the abstract knowledge representation of natural language sentences and semantic role labeling[4]；Dan Shen and Mirella Lapata testified the validity using semantic roles in question answering system[5]. These previous approaches all apply syntactic and semantic information in their system design and have achieved satisfying results. However, as these systems are mainly targeted at single sentence processing, they seem to be inadequate in dealing with complex sentences with nested clauses or modifiers and contextual sentence information extraction. Moreover, the insufficient number of semantic roles also makes it impossible to do an elaborate analysis based on the coarse semantic labeling. These drawbacks might be a hindrance to the performance of the complex question answering systems.

Suppose there is such an information retrieval system, where the documents in each file are “segmented” into lexical-level “fragments” with sense in sentence before they are saved into database. When answering questions, the system only needs to segment the input question sentence into “fragments” of a similar kind, and then retrieve the combination of these “fragments” from the database to produce an answer corresponding to the question sentence. In other words, the system is trying to realize a transfer, in which a document is first segmented into a number of lexical-level “fragments” with sense in sentence, then the relevant information among these “fragments” is picked up and finally an original document is generated by reorganizing the “segmented” “fragments”. Moreover, these “fragments” can also be retrieved separately. Figure 1 below is an illustration of the point.

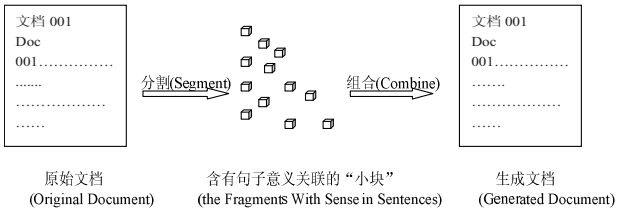



**Figure 1. The segmentation and generation of document**

Note: The “fragments” here refer to the lexical-level units with sense in sentence rather than the words in the conventional sense. In addition to the morphological information of words, the “fragments” also contain the syntactic and semantic information. For instance, when such a question as “昨天汤姆给约翰什么？”(What did Tom give John yesterday?) is input, the system can retrieve a sentence like “昨天汤姆给约翰1本书” (Tom gave John a book yesterday.), excluding the results like “约翰昨天给汤姆一个苹果” (John gave Tom an apple yesterday.) and “汤姆的哥哥昨天给约翰 1 本书”(Tom’s brother gave John a book yesterday.), though these sentences all contain “Tom”, “gave” and “John” three words.

If such an information retrieval system could be realized, the question answering system would be able to locate accurate information, generate automatically answers on the basis of the clues provided by the "fragments", and get a "Nugget" level simplified answer. We seem to have found a close equivalent to such a "fragment", namely, the Situation Unit (SU), on which our Chinese question answering system DLUT is based.

## 2. SYSTEM DESCRIPTION

What follows is an elaboration on the procedure behind the SU-based question answering system.

### 2.1 Defining Situation Unit (SU)

#### *2.1.1 Situation*

Situation semantics was originally conceived by the American logician J.Barwise and psychologist J.Perry in the early 1980s and subsequently developed by K.Devlin. "Situation" in this paper is cited in the sense defined: "The world consists not just of objects, or of objects, properties and relations, but of objects having properties and standing in relations to one another. And there are parts of the world, clearly recognized (although not precisely individuated) in common sense and human language. These parts of the world are called situations. Events and episodes are situations in time, scenes are visually perceived situations, changes are sequences of situations, and facts are situations enriched (or polluted) by language."[7]. Masahito Kawamori applies situation semantics to the analysis of Japanese argument roles [8].

#### *2.1.2 The Situation Unit （SU）*

Every situation can be divided into some small parts by different ways. In situation semantics, these small parts are *space-time location*, *properties*, *individuals*, *relations*, *Infon*, *Profon,* etc, but we divide situations into some small parts by other ways. As the situation can be described by natural language sentences and every sentence consists of some words, we think that the every word of sentence surely embodies the partial information of situation. In order to avoid confusion with those small parts existed in situation semantics, we call such partial information of situation "Situation Unit" (SU), which is embodied by the word of sentence. A SU is expressed by a word coupled with a combination of the information attributed to this word after syntactic and semantic analysis. Therefore, the quantity of the SU in a natural language sentence is dependent on the quantity of words of that sentence.

#### *2.1.3 The Constituents of the SU*

Each SU comprises of 10 constituents, as shown below:

**Table1. The constituent of the SU**

| No._Record | Unit_Type | Sytactical_Constituent_Code | Semantic_Code | Word |
|---|---|---|---|---|
| No._Document | No._Sentence | No._Clause | Hierachical_Code | Sytactical_Constituent_Frame |

An example is given to illustrate the point. The sentence "钢琴(Piano)进入(Enters)中国(China)城市(City)寻常(Ordinary)百姓家(Family)" consists of 6 words, whose corresponding SU is listed below:

**Table 2. The sample with 6 SUs**

1、000010 EVENT 031 Hj64 进入(Enters) XIN_CMN_20020101.0001#006 0*0 031011041
2、000011 EVENT 011 Bp13 钢琴(Piano) XIN_CMN_20020101.0001#006 0*0 031011041
3、000012 EVENT 041 Wr01 百姓家(Family) XIN_CMN_20020101.0001#006 0*0 031011041
4、000013 ATTRIBUTE 331 Cb25 城市(City) XIN_CMN_20020101.0001#006 0*0-041 331511
5、000014 ATTRIBUTE 511 Ed04 寻常(Ordinary) XIN_CMN_20020101.0001#006 0*0-041 331511
6、000015 ATTRIBUTE 421 Di02 中国（China）XIN_CMN_20020101.0001#006 0*0-041-331 421

*The Term of Constituents of the SU*

**No._Record:** the only record number corresponding to each specific SU in the database.

**Unit_Type:** If the word of a SU acts as the subject, predicate, object of a sentence, or as an adverbial word modifying a verb, the SU is classified into the EVENT unit type; if it modifies the subject or object of a sentence/clause, the SU is classified into an ATTRIBUTE unit type.

**Syntactical_Constituent_Code:** Different code is assigned according to the role the SU word plays in a sentence/clause. It might be the subject, predicate, object, adverbial, complement, or the appositive of a subject or object. For example, "01" represents the subject, "02" represents the appositive of a subject, "03" represents the predicate. As one sentence might contain several the same syntactic elements, it is necessary to indicate the order of them through marking. For example, if a subject marked as "01" is the second subject in a sentence, then its Syntactical_Constituent_Code would be "012".

**Semantic_Code**: The SU is semantically marked according to the *《同义词词林》* (*Chinese Synonym Dictionary*), in which only the first 4 marking numbers are adopted, so the word "国家" (country) and "人"(people) would be marked as "Dn04" and "Af01" respectively.

**Word**: The word in sentence or clause, such as "国家" (country) .

**No._Document**: the serial number of the document where the SU word appears. So, "XINHUA_CMN_20020102.0100" would indicate the 100th document of 《新华报》(Xinhua News) (Chinese version) published on January 2, 2002.

**No._Sentence**: the marking number of the sentence in which the SU word occurs, hence, "024" refers to the 24th sentence.

**No._Clause**: the marking number of the clause in which the SU word occurs, hence, "002" represents the second clause.

**Hierachical_Code:** As a situation sentence may contain several clauses, and the syntactic element may have modifiers, a SU may have been inherited from its paternal hierarchy, yet it has its own sub-categories as well. Take the nested structure marking of a SU "0-011-031" for example. If the Unit_Type of the SU is "ATTRIBUTE", the attributive of the first subject ("011") of the root sentence ("0") is the No.031 of the categories of the "ATTRIBUTE". Otherwise, if the Unit_Type of the SU is "EVENT", the attributive of the first subject ("011") of the root sentence ("0") is a clause marked by the first predicate ("031").

**Syntactical_Constituent_Frame**: The major syntactic element in a situation sentence is expressed as a frame. Every three characters symbolize one syntactic element. For instance, "011031041" indicates the first subject "011", the first predicate "031" and the first object "041".

### *2.1.4 The features of the SU are as follows:*

I. The answer to question can be retrieved through matching the SUs in the questions sentence with those in the text of corpus. Flexibility is allowed in the matching process, which means that it might get the rest of constituents of SU by matching one of constituents of SU, or matching a combination of multiple constituents of SU. The matching a combined multiple constituents of SU might improve accuracy.

II. As Semantic_Code constituent of SUs is directly related to the coding in accordance with the 《同义词词林》 (*Chinese Synonym Dictionary*) which includes 95 major categories and 1428 medium categories rather than about 20 categories of semantic role labeling [9], the classification is more specific and the treatment is more exact. In addition, this can avoid the errors which are likely to appear in semantic role labeling.

III. Some semantically similar synonyms can be matched at one time by matching Semantic_Code of SU, which is more time-efficient than the conventional keyword extension approach which matches synonyms one by one.

IV. As Semantic_Code of SU is given from the word in sentence rather than an isolated word, it is possible to match a word according to the exclusive meaning appointed in the sentence by matching Semantic_Code of SU. This is very helpful to ambiguous word matching.

V. Through adjusting the number of the No._Sentence or No._Clause of SU by "+1" or "-1", the relevant SU information about the preceding or the following of the given sentence or clause may be accessed quickly, which makes it possible to get the contextual information of sentence by cross-sentence access.

VI. It is possible to know whether a syntactic element is modified by an attributive word or a clause by the combination of the Unit_Type and the Hierarchical_Code of SU, which is extremely essential for getting the structural clues of a complex sentence with hierarchical clauses. This might be used to analyze some complex sentences.

VII. The original answer sentence "segmented" into SUs can be generated by "reorganizing" the SUs of the answer sentence and can be simplified by some rules. This means that the corresponding answer sentence to questions can be accessed just through SUs while the text corpus is not used for reference, which might save the space for corpus storage and avoid analyzing repetitively the same text of corpus to get syntactic and semantic information..

VIII. The drawback of SU is that it needs more storage space than a single word, even more when the corresponding indexes are set, though the problem may become less prominent with the advance of the storage techniques and the decline of the cost, coupled with some data compressing techniques.

## 2.2 The Procedure of the SU-based Question Answering System

The working procedure is mainly divided into three sections: extracting the SUs from the texts of corpus; extracting the SUs from the question sentences; matching SUs and the answer output, as illustrated below in figure 2:

### *2.2.1 Extracting the SUs from the texts of corpus*

#### *2.2.1.1 Analyzing the given text corpus syntactically and semantically*

To perform syntactic and semantic analysis of the text of corpus, the Chinese language processing free software package LTP 2.0 from the Information Retrieval Center of HIT has been applied [10]. The syntactic and semantic analysis result for "钢琴进入中国城市寻常百姓家"(Piano enters the ordinary family in China city ) is shown as below:

**Table 3. The sample for syntactic and semantic analysis result**

| *No._Document:* | *XIN_CMN_20020101.0001→6:* | | | | | |
|---|---|---|---|---|---|---|
| *Word:* | *钢琴(piano)* | *进入(enters)* | *中国(China)* | *城市(city)* | *寻常(ordinary)* | *百姓家(family)* |
| *POS:* | *N* | *v* | *ns* | *n* | *a* | *n* |
| *Dependency_Dep.:* | *1* | *-1* | *3* | *5* | *5* | *1* |
| *Dependency_ReL.:* | *SBV* | *HED* | *ATT* | *ATT* | *ATT* | *VOB* |
| *Semantic_Code:* | *Bp1* | *Hj64* | *Di02* | *Cb25* | *Ed04* | *-1* |

#### *2.2.1.2 Transferring the syntactic and semantic analysis results into the syntactic and semantic frame*

As it is still not an easy task to extract SU directly from the syntactic and semantic analysis results, a syntactic and semantic frame needs to be generated so that an SU can be transferred through this syntactic and semantic frame. The syntactic and semantic frame for "钢琴进入中国城市寻常百姓家."(Piano enters the ordinary family in China City.) is shown as below:

**Table 4. The sample for syntactic and semantic frame**

| 编号 number | 单元0 column0 | 单元1 column1 | 单元2 column2 | 单元3 column3 | 单元4 column4 |
|---|---|---|---|---|---|
| 0 | 层级 Hierachy | 事件 EVENT | 1谓语-1 1 Predict-1 | 1主语-1 1 Subject-1 | 1间接宾语-1 1 Indir-Object-1 |
| 0 | 0*0 | 事件 EVENT | 1 进入 Hj64 (Enters) | 0 钢琴 Bp13 (Piano) | 百姓家/Wn01 (Family) |
| 1 | 层级 Hierachy | 属性 ATTRIBUTE | 名词/Wn-1 (Noun) | 空间/Cb-1 (Space) | 性质/Ed-1 (Characteristics) |
| 1 | 0*0-041 | 属性 ATTRIBUTE | 5 百姓家/Wn01 (Family) | 3 城市 Cb25 (City) | 4 寻常 Ed04 (Oridinary) |
| 2 | 层级 Hierachy | 属性 ATTRIBUTE | 空间/Cb-1 (Space) | 社会政法/Di-1 (Social Country) | |
| 2 | 0*0-041-331 | 属性 ATTRIBUTE | 3 城市 Cb25 (City) | 2 中国 Di02 (China) | |

Note: the Semantic_Code of the SU for "百姓家 (Family)" is "-1", indicating there is no corresponding Semantic_Code. In order to do post-processing, the system will automatically produce a Semantic_Code "Wn01" for it according to its POS "n" (noun).

#### *2.2.1.3 Extracting the SU directly from syntactic and semantic frame*

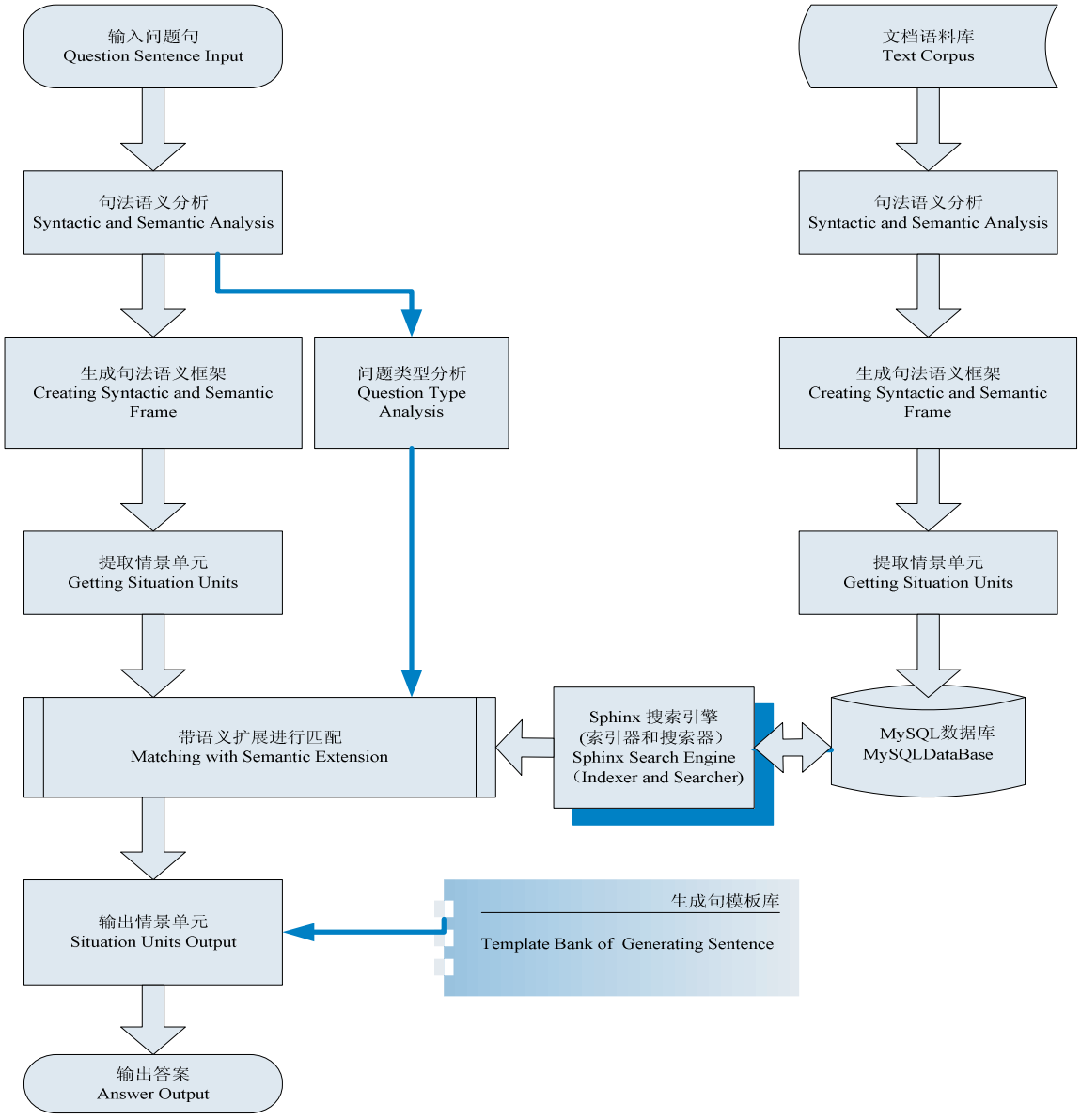
Extract "EVENT" or "ATTRIBUTE" according to "Unit_Type"; extract the Hierarchical_Code of the SU according to the inherited relations of the modifiers; extract the rest corresponding information in turn by referring to the Semantic_Code, Words, No._Document, etc. The extraction of the SU is realized by fully extracting all of constituent information.

#### *2.2.1.4 Writing the SUs of the Text of Corpus into a Database*

Every constituent in the SU can be written into the database for storage. The data may be compressed according to the different situations. For instance, "EVENT/ATTRIBUTE" can be represented by a single binary code "0/1", Words by their index in the vocabulary list, No._Document information by the index of their document number, etc. Compressed storage can reduce the cost of the data storing space.

#### *2.2.1.5 Using the Search Engine Sphinx SE and the Database MySQL to Set Indexes and Speed Retrieval*

In order to increase the retrieval speed, some indexes of the essential information needs to be set, such as Semantic_Code, Syntactical_Constituent_Code, Word, etc. Our system DLUT adopts the high quality full text search engine Sphinx SE by Andrew Aksyonoff [11], as it is compatible with the database MySQL, which is used for SU data storage in our system.




基于情景单元的问答系统的工作流程
The Procedure of Question Answering System Based on Situation Unit

**Figure 2: The Procedure of the SU-based Question Answering System**

### 2.2.2 Extracting the SUs of the Input Questions

Basically, the procedure to extract the SU of the input questions is similar to the procedure to extract the SU from the text of corpus except involving one more step, the question type analysis, an essential component of the question answering system. Our system applies the template matching approach in which the question type analysis is conducted by matching it with the question type template bank. If necessary, the key words of the questions are extracted and extended by Semantic_Code. For example, if "Di02" is used for "国家" (country) as in "哪个国家" (which country) , it can cover many words which are relevant to "国家"(country).

For example, the question "ACLIA2-CS-0009: 2008 奥运会在哪里举办？"(where will Olympic Games 2008 be held?), because it is known that such question's type is "LOCATION" through the pattern matching method, the SU corresponding to the question word of "where" will be not used, but the system will provide the Semantic_Code "Di02/Di03" of "LOCATION" for matching. The SUs of this question sentence are shown as below:

**Table 5. The SUs of question sentence "2008 奥运会在哪里举办？"(where will Olympic Games 2008 be held?）**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| EVENT | 031 | Hc05 | 举办 (Held) | INPUT-0 | 0*0 | 031011 | |
| EVENT | 011 | Di23 | 奥运会 (Olympic Games) | INPUT-0 | 0*0 | 031011 | |
| ATTRIBUTE | 471 | Dn04 | 2008 | INPUT-0 | 0*0-011 | 471 | |
| ? | ? | Di02/Di03 | ? | ? | ? | ? | ? |

### 2.2.3 Matching the SUs and the Answer Output

The system is able to get answers by matching the SUs contained in a question sentence with those in text of corpus in the database. The specific matching conditions are determined mainly by the question type and the matching accuracy requirement. The following is an example on how to get the answer to the question "ACLIA2-CS-0009: 2008 奥运会在哪里举办？"(where will Olympic Games 2008 be held).

#### 2.2.3.1 Matching the SUs of the Question Sentence in the Database

a. As the Syntactical_Constituent_Code "031" of the SU embodied by the word "hold" is as the predicate, select the Semantic_Code "Hc05" of the SU to match in the database, the partial matched results are shown as below:

**Table 6. The SUs matched with Semantic_Code "Hc05"**

```
.....
5326580 事件 461 Hc05 举办 XIN_CMN_20020501.0111#004 4*0 031081082461
5326581 事件 031 Hc05 举办 XIN_CMN_20020501.0111#004 4*0-461 031041
5327411 属性 A41 Hc05 举办 XIN_CMN_20020501.0113#013 1*0-911 A41
5327412 事件 031 Hc05 举办 XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
5332522 事件 031 Hc05 举办 XIN_CMN_20020501.0135#002 2*0 031081082083041061
5334412 属性 A41 Hc05 举办 XIN_CMN_20020501.0142#015 0*0-081-331-421-911 A41
.....
```

b. As the Syntactical_constituent_Code "011" of the SU embodied by the word "Olympic Games" is as the subject, the word "Olympic Games" of the SU is selected to match in the database, the partial matched results are shown as below:

**Table 7. The SUs matched with Word "奥运会"**

```
.....
5321829 属性 681 Hh07 奥运会 XIN_CMN_20020501.0093#005 1*0-011-911-A41-081-331 681951
5327160 事件 041 Hh07 奥运会 XIN_CMN_20020501.0113#002 0*0-431 031041
5327406 属性 681 Hh07 奥运会 XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327429 属性 681 Hh07 奥运会 XIN_CMN_20020501.0113#014 0*0-041-911-331 681
5327488 事件 041 Hh07 奥运会 XIN_CMN_20020501.0113#015 4*0-131-A41 031001041
.....
```

c. As the Unit_Type of the SU embodied by the word "2008" is "ATTRIBUTE" , the Hierarchical_Code is "0-011" (the first subject "011" of the root sentence "0"), and the Semantic_Code of the SU is "Dn04" which is a numeral and as the attribute, the Word "2008" of the SU is selected to match in the database, the partial matched results are shown as below:

**Table 8. The SUs matched with Word "2008"**

```
.....
4629061 属性 471 Dn04 2008 XIN_CMN_20020411.0102#006 0*0-461-011 471422423
4785299 属性 471 Dn04 2008 XIN_CMN_20020415.0010#023 1*0-041-911-681 471421
5025156 属性 472 Dn04 2008 XIN_CMN_20020419.0127#004 0*0-041 472
5327409 属性 471 Dn04 2008 XIN_CMN_20020501.0113#013 1*0 681881911471A01
.....
```

d. As the Semantic_Code *"Di03" of the SU* is embodied by the question word "where" of "LOCATION" question type, the Semantic_Code "Cb25" is selected to match in the database, the partial matched results are shown as below:

**Table 9. The SUs matched with Semantic_Code "Di03"**

```
.....
5327323 事件 011 Di03 北京 XIN_CMN_20020501.0113#010 3*0-011-911-A41 031011081041
5327337 属性 421 Di03 北京 XIN_CMN_20020501.0113#011 1*0-041-911 421511
5327413 事件 011 Di03 北京 XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
5327431 事件 011 Di03 北京 XIN_CMN_20020501.0113#014 0*0-011-A41 031011081082083
5327498 属性 421 Di03 北京 XIN_CMN_20020501.0113#016 1*0-011 421
.....
```

#### 2.2.3.2 Grouping the Matched SUs Based on the Same No._Document, No._Sentence and No._Clause

In the above mentioned matching, the matched results are obtained by matching a certain constituent of the SU of the question sentence, and they may scatter in different document, sentence or clause but we only need the matched results that all SUs of the question sentence are contained in the same sentence or clause of the same document, so it is necessary to group the matched SUs based on the No._Document, No._Sentence and No._Clause , and the matched SUs can be classified into several groups.

As for that whether each group is in a descending order according to the appointed weight of the SUs matched with those of the question sentence wholly or partly, in this participation, the system only takes the groups wholly containing SUs of the question sentence as the corresponding candidate answer groups regardless of the Syntactical_Constituent_Code of the SUs matching. For example, the two candidate groups corresponding

to the same No._Document, No._Sentence and No._Clause to the question "ACLIA2-CS-0009: 2008 奥运会在哪里举办？" ( where will Olympic Games 2008 be held ? ) are shown as below:

**Table 10. The matched SUs grouped according to the same No._Document, No._Sentence and No._Clause**

the groups of the matched SUs

5327406 属性681 Hh07 奥运会XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327409 属性471 Dn04 2008 XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327412 事件031 Hc05 举办XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
5327413 事件011 Di03 北京XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
==========
13036572 事件031 Hc05 举办 XIN_CMN_20021013.0049#031 0*0 031011041061
13036573 事件011 Di03 北京XIN_CMN_20021013.0049#031 0*0 031011041061
13036574 事件041 Hh07 奥运会XIN_CMN_20021013.0049#031 0*0 031011041061
13036576 属性471 Dn04 2008年XIN_CMN_20021013.0049#031 0*0-041 471

GROUP 0011

GROUP 0012

### 2.2.3.3 Getting the All SUs from the database according to the Same No._Document ,No._Sentence and No._Clause

As the SUs of a sentence are grouped in the same group, all SUs of the sentence can be got from database by retrieving the No._Document, No._Sentence and No._Clause of the group. For example, the two groups of all SUs are shown as below:

**Table 11. The all SUs retrieved from the database according to the same No._Document, No._Sentence and No._Clause**

the groups of all SUs

5327406 属性681 Hh07 奥运会XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327408 属性911 Kd01 的XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327409 属性471 Dn04 2008 XIN_CMN_20020501.0113#013 1*0 681881911471A01
5327411 属性A41 Hc05 举办XIN_CMN_20020501.0113#013 1*0-911-A41
5327412 事件031 Hc05 举办XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
5327413 事件011 Di03 北京XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
5327414 事件081 Ka13 将XIN_CMN_20020501.0113#013 1*0-911-A41 031011081
==========
13036572 事件031 Hc05 举办 XIN_CMN_20021013.0049#031 0*0 031011041061
13036573 事件011 Di03 北京XIN_CMN_20021013.0049#031 0*0 031011041061
13036574 事件041 Hh07 奥运会XIN_CMN_20021013.0049#031 0*0 031011041061
13036575 事件011 Wh01 XIN_CMN_20021013.0049#031 0*0 031011041061
13036576 属性471 Dn04 2008年XIN_CMN_20021013.0049#031 0*0-041 471
13036639 事件031 Je04 让XIN_CMN_20021013.0049#031 0*0-461 031041061
13036640 事件041 Ac01 人XIN_CMN_20021013.0049#031 0*0-461 031041061
13036641 事件061 Ga14 流连XIN_CMN_20021013.0049#031 0*0-461 031041061

GROUP 0011

GROUP 0012

### 2.2.3.4 Generatting the Sytactical_Constituent_Code Chain

In order to generate an answer sentence the system need to generate a Sytactical_Constituent_Code chain by the Unit_Type, Sytactical_Constituent_Code and Hierachical_Code of the all SUs of a group. For example, the three Sytactical_Constituent_Code chains of the Group 0011("XIN_CMN_20020501.0113 # 013") are shown as below:

*ATTRIBUTE 1*0 : 681+911+471*

*ATTRIBUTE 1*0-911 : A41*

*ATTRIBUTE 1*0-911-A41: 031 + 011 + 081*

### 2.2.3.5 Generating the Answer Sentence by Matching Sytactical_Constituent_Code Chains in Template Bank of Generating Sentence

An answer sentence is generated according to the word order which is given by matching Sytactical_Constituent_Code chains in the pre-established template bank for generating sentence. For example, the whole answer sentence generated by organizing the Words of SUs with the word orders matched by the Sytactical_Constituent_Code chain from template bank. The details are shown as below:

**Sytactical_Constituent_Code Chain**: **Word Order**

*ATTRIBUTE 681-911-471: [911] [471] [681]*
*→ "的"(of) + " 2008" + "奥运会"*

*ATTRIBUTE A41:no content ( only link modifiers)*

*EVENT 031-011-081: [011] [081] [031]*
*→ 北京(Peking) 将(will) 举办(held)*

***the whole sentence generated:北京 将 举办+ 的 2008 奥运会***
***→ 北京将举办的 2008 奥运会***

Notes: The above describes how to generate answer sentences by matching Sytactical_Constituent_Code chains. However, in this evaluation, our system directly called the whole answer sentences from the text of corpus rather than generated answer sentence by matching templates in template bank of generating sentence during to time limit and unfinished template bank.

## 3. EVALUATION RESULTS

The average F3 score of our system in this evaluation is F3 Score = 0.1954 (Official); F3 Score = 0.3154(Auto).

### 3.1 Scores of Various Question Types

**Table 12. The average F3 Score, Recall, Precision of various question types, Average quantity of candidates**

| Answer Type | BIOGRAPHY | DATE | DEFINITION | EVENT | LOCATION | ORGNIZATION | PERSON | RELATIONSHIP | WHY | **ALL** |
|---|---|---|---|---|---|---|---|---|---|---|
| Average F3 Score | 0.5014 | 0.0717 | 0.3164 | 0.1638 | 0.1208 | 0.2443 | 0.1559 | 0.2078 | 0.0483 | **0.1954** |
| Average Recall | 0.8766 | 0.6 | 0.6821 | 0.2904 | 0.8 | 0.6 | 0.8 | 0.3208 | 0.0561 | **0.4332** |
| Average Precision | 0.1147 | 0.0095 | 0.0752 | 0.0793 | 0.0156 | 0.0674 | 0.0222 | 0.0672 | 0.0271 | **0.0595** |
| Average quantity of candidates | 18.6 | 20.8 | 22.4 | 12.1 | 19.6 | 9 | 8.6 | 13.9 | 5.6 | **13.32** |

### 3.2 F3 Score, Recall, Precision of various Question Types according to Type Distribution

BIOGRAPHY：1-10；DATE：11-15；DEFINITION：16-25；EVENT：26-45；LOCATION：46-50；ORGNIZATION：51-55；PERSON：56-60；RELATIONSHIP：61-80；WHY：81-100；

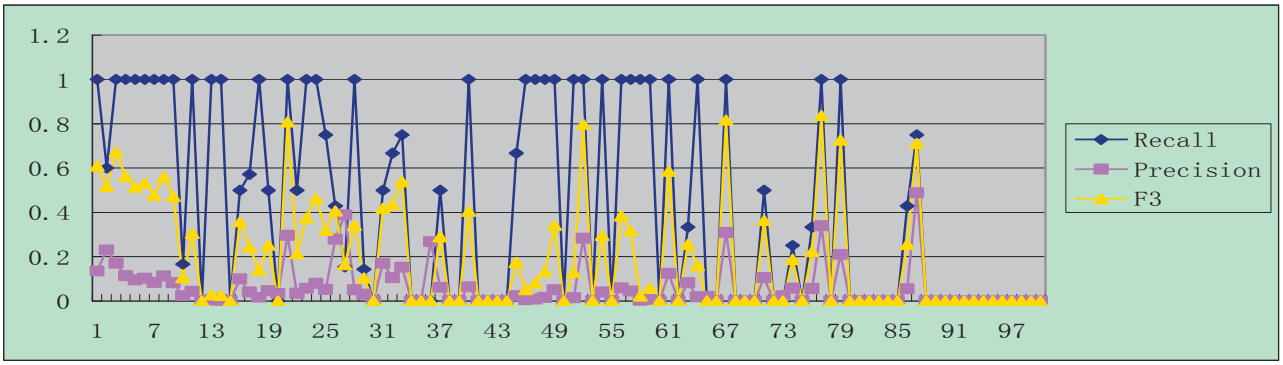



Figure3. **The Recall, Precision, F3 of Various Question Types According to Type Distribution**

The evaluation results indicate that the average recall score is ≥0.75 for the six types of questions, BIOGRAPHY、DATE、DEFINITION、LOCATION、ORGNIZATION、PERSON. As our system does not perform cross-sentence access and text structure analysis, questions like EVENT、RELATIONSHIP and WHY are not taken into account. The recall scores of the six question types are as follows:

3 of them ≥0.8; 3 of them ≥0.6, among which one DATE type question might not have a wrong answer if the system record had not been wrong. If this is taken into account, the average recall score would reach 0.7598.

Moreover, it can be seen from figure3. that 33 of 100 questions, Recall = 1，indicating one-third of the question sentences have successfully recalled their answers.

## 4. ERROR ANALYSIS

Comparing our results with the gold-standard answers of the NTCIR-8, the errors can be categorized as follows:

i. The parser applied in this evaluation, the Chinese language processing package LTP2.0 (June 2009 version) from the Information Retrieval Center of HIT still generates some syntactic analysis errors. For example, because "福克斯" in "ACLIA2-CS-0037：加菲猫和 20 世纪福克斯公司的关系是什么？" is mistakenly processed in syntactic analysis, the result of this question is completely wrong.

ii. There are some question that needs cross-sentence analysis to get the right answer, as shown by the answer to the question "ACLIA2-CS-0089 ：严凯泰与中国之间有何渊源？". The words "中国" and "严凯泰" in the answer ".....旨在表彰中国内地及港澳台地区青年经济精英的评选活动日前揭晓。神州数码有限公司总裁郭为、台湾裕隆集团执行长严凯泰.....。" belong to two clauses. Although our system has the ability for cross-sentence analysis, it did not perform this function in this evaluation, leading to analysis errors to the questions that need such a function.

iii. Further analysis is needed to some question types to generate a right answer. For instance, "ACLIA2-CS-0032：刘翔和约翰逊什么关系" belongs to the last few candidate documents in "XIN_CMN_20050325". Our system fails to perform sufficient analysis in this aspect, so too many candidate answer sentences are generated. When only the first 30 candidate sentences are used as answer, some valid answers are excluded. As a result, the recall score for these types of questions are very low.

iv. Some question sentences need to be analyzed from the perspective of a paragraph. For instance, the answer for "ACLIA2-CS-0027：蜘蛛为什么能在天花板上行走？" has to be generated through text structure analysis, which our system does not perform, finally leading to result errors. This is mainly shown through the "WHY" question types. Among the 20 "WHY" questions, the F3 for 18 of them is 0.

v. As our system extracts the corresponding sentences directly instead of simplifying them for the final answer, the final evaluation results are very much affected. The figure 3 shows that for many question types, the Recall = 1，but as the Precision is very low, F3 is often very low. A case in point is "ACLIA2-CS-0041 ： 请问谁是 2004 年美国民主党总统候选人", whose Recall = 1 , F3 =0.0223.

## 5. CONCLUSION

This paper describes the question answering system DLUT for the evaluation of NTCIR-8 CS-CS SubTask. Owing to the time limit, the rich resources of the SU are only used partly in our system. So the preliminary conclusion to be reached so far through our present experiment is that a better recall score may be got by an SU-based question answering system in view of the average recall score ≥0.75 for the six question types BIOGRAPHY, DATE, DEFINITION, LOCATION, ORGNIZATION, PERSON.

The errors of word segmentation and named entity recognition may result in analysis mistakes. The performance of the parser decides the quality of the whole system. If answers output in the form of whole sentences instead of simplified answers may greatly affect the final F3 Score. The extracting answer without cross-sentence access and text structure analysis has a big impact on the results. All this is to be improved in our future system.

## ACKNOWLEDGEMENT

Our gratitude goes to the Information Retrieval Center of HIT for their Chinese language processing package LTP2.0.

## REFERENCES

[1] Boris Katz, Gary Borchardt, et al. 2005.Syntactic and Semantic Decomposition Strategies for Question Answering from Multiple Resources. Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering, 35-41, July, 2005.

[2] Hyeon-Jin Kim, Ji-Hyun Wang, et al. 2005. LF Based Answer Indexing Method for Encyclopedia Question-Answering System. AIRS 2005, LNCS 3689, pp. 679–684.

[3] Barney Pell. 2007. Natural Language and the Semantic Web. ISWC 2007.

[4] Daniel G. Bobrow, Bob Cheslow, Cleo Condoravdi, et al. 2007.PARC's Bridge and Question Answering System. Proceedings of the GEAF2007 Workshop.

[5] Dan Shen, Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational.

[6] Jon. Barwise, John Perry. 1983. Situations and Attitudes , Bradford Books, The MIT Press.

[7] Keith Devlin. 2008. Situation theory and situation semantics.Handbook of the History of Logic, Volume 7, Elsevier, pp. 601-664.

[8] Masahito Kawamori. 1991. A Situation Semantic Approach to Argument-roles in Japanese. THE SIXTH JAPANESE-KOREAN JOINT CONFERENCE ON FORMAL LINGUISTICS.

[9] Daniel Gildea, Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245–288.

[10] http://ir.hit.edu.cn/demo/ltp/

[11] http://sphinxsearch.com/