NiCT/ATR in NTCIR-7 CCLQA Track: Answering Complex Cross-lingual Questions

Youzheng Wu†‡ Wenliang Chen† Hideki Kashioka†‡
† National Institute of Information and Communications Technology (NiCT)
‡ ATR Spoken Language Communication Research Labs
2-2-2 Hikaridai “Keihanna Science City”, Kyoto 619-0288, Japan
{youzheng.wu,chenwl,hideki.kashioka}@nict.go.jp

Abstract

This paper describes our complex cross-lingual question answering (CCLQA) system for NTCIR-7 ACLIA track. To answer complex questions such as events, biographies, definitions, and relations, we designed two models, i.e., the centroid-vector model and the SVM-based model. In the official evaluation of the NTCIR-7 CCLQA track, our SVM-based model achieved 22.11% F-score in the English-Chinese cross-lingual task, the highest score among all participants' systems, and 23.16% F-score in the Chinese-Chinese monolingual task. In the automatic evaluation, the F-scores of the SVM-based model and the centroid-vector model in the English-Chinese task are 27.24%, and 24.55%,respectively. In the Chinese-Chinese task, the two models achieved 28.30%, and 24.78% F-scores.

Keywords: *Complex Question Answering (QA), Cross-lingual QA, Chinese, English.*

1 Introduction

In this year, NTCIR cross-lingual question answering (CLQA) track shifts focus from factoid CLQA to complex CLQA that is referred to as CCLQA. This task is novel because no evaluation at NTCIR, TREC or CLEF has evaluated cross-lingual QA on complex questions. Although the main focus is cross-lingual complex QA, the organizers also intend to accept monolingual complex QA runs. The CCLQA task is asking complex questions in English and getting answers in the given Chinese (Simplified, Traditional) or Japanese corpora. The complex questions involved include four types of questions, i.e., events (`List major events in formation of European Union.`), biographies (`Who is Howard Dean?`), definitions (`What are stem cells?`), and relations (`What is the relationship between Saddam Hussein and Jacques Chirac?`). The corpora used are showed in Table 1.

Table 1. The Corpora Used in CCLQA.

Language	Corpus Name	Span
Simplified Chinese	Xinhua	1998-2001
	Lianhe Zaobao	1998-2001
Traditional Chinese	CIRB020 & CIRB040	1998-2001
Japanese	Mainichi Newspaper	1998-2001

This year is our first time to participate NTCIR tracks. We just selected the Simplified Chinese-related CLQA tasks, i.e., English - Simplified Chinese, and Simplified Chinese - Simplified Chinese tasks. In the following sections, unless specified, otherwise, Chinese denotes Simplified Chinese; EN and CS are the abbreviations of English and Simplified Chinese.

2 NiCT/ATR CCLQA System

We implemented two models for the CCLQA tasks. One is the centroid-vector model, the other is the SVM-based model. The main ideas between the two models are completely different. The architectures are illustrated in Figure 1, where, the centroid-vector model is marked with broken lines, and the SVM-based model is marked with point lines.

Figure 1 shows that three modules are shared by the two designed models. Given the same candidate answers $\{s_i|i = 1, 2, ..., n\}$ to test question, the two models, however, use different approaches to select some of them as answers.

In Section 3, we describe the common modules shared by the centroid-vector model and the SVM-based model. Their specific modules are introduced in Section 4 and Section 5. Section 6 presents the results of our systems for the CCLQA formal runs. Section 7 summarizes our work.

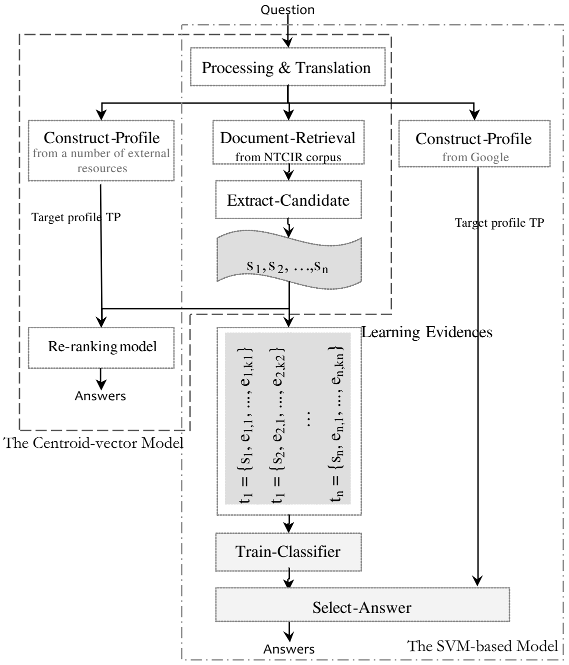

Figure 1. The architecture of our system

3 Common Modules

This section describes the common modules shared by the two models, i.e., the $processing$ & $translation$,the $document\ retrieval$, and the $extract\ candidate$ modules.

3.1 Processing & Translation

The CCLQA track provides four fields such as EN QUESTION, CS QUESTION, EN NARRATIVE, and CS NARRATIVE for each test question. In the EN-CS task, only EN QUESTION and EN NARRATIVE can be used. Similarly, the CS-CS task can only use CS QUESTION and CS NARRATIVE. Note that, our system just used QUESTION field, no information from NARRATIVE are incorporated. The following is an example.

```
<TOPIC ID="ACLIA1-CS-T97">
    <QUESTION LANG="EN"><What is the relationship
    between Guo Jingjing and diving.></QUESTION>
    <QUESTION LANG="CS"><郭晶晶和跳水有什么关
    系?></QUESTION>
    <NARRATIVE LANG="EN"><I would like to know
    the relationship between Guo Jingjing and
    diving.></NARRATIVE>
    <NARRATIVE LANG="CS"><我想知道郭晶晶和跳水的关
    系。></NARRATIVE>
</TOPIC>
```

In the CS-CS task, the $processing$ & $translation$ module includes three steps: 1. extracting Chinese question target from CS QUESTION; 2. identifying type of question; 3. Chinese word segmentation and part-of-speech tagging. For the above example, the outputs of the three steps are "郭晶晶和跳水", Relation question, and "郭/nr 晶晶/nr 和/c 跳水/vn" respectively.

In the EN-CS task, this module consists of four steps: 1. extracting question target from EN QUESTION, referred to as English question target; 2. identifying type of question; 3. translating English question target into Chinese, the translation result is also called Chinese question target. 4. Chinese word segmentation and part-of-speech tagging. For the above example, the output of the four steps are "Guo Jingjing and diving", Relation question, "郭晶晶和跳水", and "郭/nr 晶晶/nr 和/c 跳水/vn" respectively.

To realize step 1 and 2 in the EN-CS and CS-CS tasks, we manually formulated a set of patterns for each type of questions from the dry-run data set. For instance, the pattern "*what is the relation between X*" indicates that the type of question is Relation, and *X* is question target.

To realize step 3 in the EN-CS task, we employ two translation engines such as Google (`http://translate.google.com`) and Yahoo (`http://fanyi.cn.yahoo.com/`). The combination of the two engines can compensate for translation errors generated by either of them. For example, the English question target "*Nobel Prize*" is wrongly translated into "诺贝尔文学奖" by Google, while Yahoo generates the correct translation "诺贝尔奖".

To realize step 4 in the EN-CS task or step 3 in the CS-CS task, we adopt the tool developed by [8] to segment Chinese documents and tag part-of-speeches. Please refer to `http://icl.pku.edu.cn/icl_groups/corpus/addition.htm` for the part-of-speech tags.

3.2 Document Retrieval

We use Indri toolkit (`http://www.lemurproject.org`), an open-source toolkit widely used in IR community, to build an index of NTCIR Chinese corpus, and retrieve the relevant documents to the test questions. In the stage of building Indri index, Chinese corpus is first segmented into words by [8].

To retrieve the most relevant documents from the index of NTCIR corpus, the $document\text{-}retrieval$ formulates Indri queries using Indri query language that allows complex phrase matching, synonyms, weighted expressions, Boolean filtering, etc. For example, the Indri query for the question "*What is Moore's Law?*" is as follows.

```
1  #combine (
2   #weight(
3     0.5 #2(穆 尔 定律)
4     0.4 #1(Moore Law)
5     0.1 #combine(穆 尔 定律 Moore 法律)))
```

where, Line 3 is composed from the translation of Google. For biography and definition questions in the EN-CS task, words in English target are also used as Indri query, which form Line 4. Line 5 is a combination of words from Google and Yahoo translation engines. The values of 0.5, 0.4, and

0.1 are the empirical weights assigned to the retrieval expressions.

The output of this module is the 500 documents most relevant to the test questions.

3.3 Extract Candidate

The *extract-candidate* module splits the retrieved documents into sentences, applies certain heuristic rules to generate sentences as candidate answers, and tags parts-of-speeches using [8] for each candidate. The heuristic rules can be summarized as follows.

- For *biography* questions, we require a candidate answer to contain the first and last names of Chinese or English question target, and the distance between the first name and the last name is less than 3.
- For *definition* questions, we require a candidate answer to contain the exact string of the Chinese or English question target.
- For *relationship* and *event* questions, we rank sentences according to the percentage of the words in Chinese question target they contained, and select top 300 sentences as candidate answers.

The output of this module is a set of candidate answers $\{s_i|i = 1, 2, ..., n\}$ to each question, n is the number of candidates, its maximum value is 300.

4 The Centroid-vector Model

The centroid-vector model is based on the dependent assumption that the words in answers are statistically dependent on its question target. These words consists of a centroid vector, referred to as target profile. After obtaining target profile, the model ranks the candidate answers $\{s_i|i = 1, 2, ..., n\}$ provided by the $extract\ candidate$ module based on their similarity to the target profile. The similarity is $TF \times IDF$ score that is calculated in the framework of vector space model. We thus designed two modules to realize this goal.

4.1 Construct Profile

To construct an accurate profile, we exploit five external resources, i.e., $Wikipedia$ (http://zh.wikipedia.org/), $Google\ definition$, $Baidu\ Baike$ (http://baike.baidu.com/), $Google\ News$ (http://news.google.cn/), and $Google$ (http://www.google.com), to collect profile corpus for each question.

$Wikipedia$ is a Web-based, free-content encyclopedia. This resource has been employed by many QA systems as a source of knowledge and contains about 0.2 million Chinese entries.

$Google\ definition$ provides the definition for a query if it has one.We use this feature to extract whatever definitions that Google has found.

$Baidu\ Baike$ is a largest Web-based Chinese encyclopedia that is similar to $Wikipedia$, which includes about 1 million entries.

$Google\ News$ provides $1,000$ news sources updated continuously. For each question target, $Google\ News$ lists the titles and snippets of relevant articles. We download the top 100 results.

$Google$, a widely used Web search engine, is used to retrieve the relevant Web pages. We download the top 100 Google snippets, which are used as a backup resource for the above specific resources.

After obtaining the profile corpus from the above external resources, we conduct Chinese word segmentation and part-of-speeches tagging. The word-significance scores are then calculated using Equation (1).

$$score(w_j) = \delta_1 \times log\frac{c_1(w_j)}{\sum_j c_1(w_j)} + \delta_2 \times log\frac{c_2(w_j)}{\sum_j c_2(w_j)} \quad (1)$$

where, $c_1(w_j)$ denotes the frequency of word w_j in the profile corpus, $c_2(w_j)$ is the frequency of the word in NTCIR corpus, δ_1 and δ_2 are the weights.

Finally, we empirically select top 380 words with largest scores as target profile TP to question.

4.2 Re-ranking Model

Given a set of candidate answers $\{s_i|i = 1, 2, ..., n\}$ to question and the target profile TP, the re-ranking model can estimate the similarities between the candidates and the target profile using Equation (2) $\sim$ (4).

$$sim(s_i, TP) = \frac{\sum_j (wt(w_j, s_i) \times wt(w_j, TP))}{\sqrt{\sum_j wt(w_j, s_i)} \times \sqrt{\sum_j wt(w_j, TP)}} \quad (2)$$

$$wt(w_j, s_i) = TF(w_j, s_i) \times IDF(w_j) \quad (3)$$

$$IDF(w_j) = log(\frac{N}{DF(w_j)}) \quad (4)$$

where $TF(w_j, s_i)$ is the number of the occurrence of word w_j in candidate s_i, $DF(w_j)$ denotes the number of candidates that contain word w_j.

Lastly, the re-ranking module outputs the top 15 candidates with largest similarity scores as the final answer to question.

5 The SVM-based model

Apparently, the focus of the centroid-vector model is utilizing a variety of external resources to construct an accurate target profile. These resources, however, are not always available. Moreover, they do not always contribute positively to complex questions according to [4]. Thus, we present the SVM-based model that shifts the focus from mining a variety of external resources to applying sentence-expansion (SE) that involves learning evidences of candidate answers from the Web pertaining to whether or not candidates are definitions. The specific modules in the SVM-based model include the $construct\ profile$, $learning\ redundancy$, $train\ classifier$, and $select\ answer$ modules, as shown in Figure 1.

5.1 Construct Profile

Different from the $construct\ profile$ module in the centroid-vector model, which employs five external resources , this $construct\ profile$ module just uses Google to construct target profile. Thus, the SVM-based model is independent of some specific external resources. The other steps are same.

5.2 Learning Evidences

To overcome the shortcoming of target profile TP that just constructed from one single external resource such as Google in our case, the SVM-based model presents an algorithm to learn evidences for candidates. For each candidate s_i in S, the procedure of learning evidences of candidate s_i is given as follows.

$Step$-1: Extract k words from s_i to represent the topic of candidate s_i that is refer to as t_i. These words are called as topic words of s_i, labeled as TW.

$Step$-2: Combine topic words TW and the Chinese question target to compose a Web query and submit it to Google engine.

$Step$-3: Download the top 100 Google snippets.

$Step$-4: Retain those snippets $\{e_{i,1}, ..., e_{i,k_i}\}$ that contain all words of the English or Chinese question target and the TW as the Web evidences of s_i, k_i is the number of Web evidences of candidate s_i.[1]

In this procedure, the extraction of topic words plays an important role. Only the accurate topic words TW associated with the question target can capture the primary meaning of the candidate s_i and represent its topic as well as guarantee that the retained snippets are positive evidences pertaining to candidate s_i. During implementation, we consider only words with n, nr, ns, nt, nz, or t part-of-speeches (`http://icl.pku.edu.cn/icl_groups/corpus/addition.htm`) as candidates for topic words of s_i. We then calculate the distance, i.e., the number of words between the candidates of topic words to the question target and select the most nearest k terms as the final topic words. k is empirically set up to 3.

Therefore, the output of this step is a set of Web evidences $\{e_{i,1}, e_{i,2}, ..., e_{i,k_i} | i = 1, 2, ...n\}$ for candidate answers $\{s_i | i = 1, 2, ..., n\}$, as illustrated in the $learning\ evidences$ module of Figure 1.

5.3 Train Classifier

As Figure 1 illustrated, we assume that the n candidate answers $\{s_i | i = 1, 2, ..., n\}$ represent n topics $\{t_i | i = 1, 2, ..., n\}$. Because the learned Web evidences $\{e_{i,1}, e_{i,2}, ..., e_{i,k_i}\}$ of candidate s_i and the candidate s_i itself contain the same topic words and the question target, we thus consider that they express the same meaning, thus belong to the same topic t_i. Based on this assumption, we can train a n-topic classifier.

We utilize the LIBSVM toolkit [1], a multi-class probability SVM, and a one-against-one strategy for multi-class classification. The kernel of the SVM is the radical basis function. The classification features are described as follows.

$Bfull$: a feature that fires only if the question target in an instance occurs in the exact string of the target.

$Bbegin$: a feature that fires only if the question target occurs at the beginning of the training instance.

$Bpattern$: a feature that fires only if one of the predefined definition-patterns occurs. The patterns are extracted from the Wikipedia data called `extracted page abstracts for Yahoo`. These patterns are listed as follows.

<table>
<tr><th>Regular expression of patterns</th></tr>
<tr><td>DEF (*) [是|为] [一 q|一个]</td></tr>
<tr><td>DEF (*) [又 称|或 称|或 称为|或 简称|或 译|简称|旧 称|全 称为|全 名|学 名|也 称|也 称为|也 称作|也 叫|亦 称|亦 称为|亦 作|英文 名称|又 称为|又 称作|又 叫|又 名|又 译|原 称|原 为|原名|正式 名称|全 名 为|又 名|又 被 称为]</td></tr>
<tr><td>DEF (*) [是 ns 一个|位于 ns|是 位于|一个 位于|在 ns|坐落 于]</td></tr>
<tr><td>DEF (*) [是 指|指 的 是|是 由]</td></tr>
<tr><td>DEF (*) [发生|成立] [于|在] DEF (*) [是 ns|女|字]</td></tr>
<tr><td>DEF (*) [ns *家|现任 ns|本 名|生 卒]</td></tr>
<tr><td>DEF (*) [籍贯|出生于|生于|出生于|生于|祖籍] ns</td></tr>
<tr><td>DEF (*) ns [出生|人]</td></tr>
<tr><td>Legend: () is optional field; [] is obligation field; * denotes that any word can appear; | separates the elements within the obligation field, <i>q</i> and <i>ns</i> represent the part-of-speeches of quantifier and location name, respectively.</td></tr>
</table>

$Btime$: a feature that fires only if a time expression occurs. Question targets are usually involved in some historical timeline of events. Candidates containing time expressions tend to capture important events involving the target that can be selected as answers to the question based on the conclusion in [6].

Un-$overlap$: a feature that indicates the overlap of uni-grams between an instance and the target profile.

Bi-$overlap$: a feature that indicates the overlap of bi-terms between an instance and the target profile.

$Fsim$: a feature that implies a unigram-based TF × IDF similarity between an instance and the target profile. The value is computed using Equation (2) $\sim$ (4).

The values of the $Bfull$, $Bbegin$, $Bpattern$, and $Btime$ features are set to 0.5 if they fire, otherwise, the value is zero. The values of the Un-$overlap$ and Bi-$overlap$ features are computed as follows.

$$value = \frac{C(TP, S_{i,k_i})}{C(S_{i,k_i})} \tag{5}$$

where $C(TP, S_{i,k_i})$ denotes the number of unigrams or biterms in the target profile TP that occur in the training instance $S_{i,k}$, and $C(S_{i,k_i})$ denotes the total number of unigrams or bi-terms in S_{i,k_i}.

[1] If k_i is less than 10 for the Event and Relation questions, we relax the restrictions appropriately.

5.4 Select Answer

Assuming that there exists an "$ideal\ answer$" to complex question. The characteristic of this "$ideal\ answer$" is that all the predefined features fire in this "$ideal\ answer$": for instance, it contains the exact string of the question target; the target appear at the beginning; at least one pattern appear, it has the largest similarity with target profile, etc.

Assuming again that the topic of this "$ideal\ answer$" is topic t_j. The "$ideal\ answer$" belonging to topic t_j implies that the topic of this "$ideal\ answer$" is the same as that of candidate s_i. Thus, it is appropriate to select candidate s_i as the final answer to question.

Based on the above assumptions, the SVM-based model shifts selecting answers from candidates $\{s_1, s_2, ..., s_n\}$ to determining the topic of the "$ideal\ answer$". The $select\ answer$ module uses the trained SVM classifier to determine the probabilities of this "$ideal\ answer$" belonging to n topics.

The outputs of this module are $n < t_k, p_k >$ pairs, implying that the probability of the "$ideal\ answer$" belonging to topic t_k is p_k. The SVM-based model finally selects j candidate answers $\{s_k, k = 1, 2, ..., j\}$ with the largest probabilities as answers. j is set to 10 if the top 10th p_k is less than the average ($1/n$), else, j is 15.

6 Experiments

The CCLQA test set consists of 100 questions. The distribution of the test set over the types of questions are shown in Table 2. For evaluation, the CCLQA conducts both human evaluation and automatic evaluation.

Table 2. The statistic of test question set.

	Event	Biography	Definition	Relation
number	30	20	20	30

We submitted three runs for both the EN-CS task and the CS-CS task.

- RUN-1, the SVM-based model that only use only Google to construct target profiles, as described in Section 5.1.
- RUN-2, the SVM-based model. Target profiles, however, are constructed from the five external resources, same as the centroid-vector model (as described in Section 4.1).
- RUN-3, the centroid-vector model that uses the five external resources to construct target profiles, as described in Section 4.

6.1 CCLQA Track

The CCLQA track considers the first priority runs submitted by participants to be the official results, which are based on human evaluation. In our case, it is the RUN-1.

Table 3 summarizes the official results over types of questions by Recall, Precision and F-score. Note that, the value of beta in F-score is set to 3.

Table 3. The official scores of the RUN-1

	Recall (%)	Precision (%)	F-score (%)
the EN-CS task			
Event	20.70	4.8	14.54
Definition	50.13	4.2	22.16
Biography	67.04	7.2	31.58
Relation	44.78	5.6	23.35
all	43.08	5.4	**22.11**
the CS-CS task			
Event	20.87	4.3	14.30
Definition	55.38	4.9	24.15
Biography	69.40	7.9	33.76
Relation	48.15	5.6	24.29
all	45.66	5.5	**23.16**

This table demonstrates that:

- For the EN-CS task, the F-score of the RUN-1 over all test questions is 22.11%, the highest score among all participants' systems. The F-score of the next best system is 19.30% [5]. The RUN-1 in the CS-CS task achieved 23.16% F-score, this is ranked second among all systems. The F-score of the best system in the CS-CS task is 43.29% [2].
- The CCLQA track uses equation (6) to compute precision,

$$precision = \begin{cases} 1 - (L - A)/L & \text{if } L < A \\ 1 & \text{else} \end{cases} \quad (6)$$

  where L denotes the number of character-length of the system response, $A = C \times a$, a is the number of answers matched in the system response, C is character allowance per match. Using a sample from a sentence-aligned corpus, the CCLQA track estimated $C = 39$ for CS.

  However, the average number of characters of our system response is 184. As a result, the precision is relatively low.
- From this experiment, we can conclude that the performance ranking of answering these questions is: Biography > Definition > Relation > Event. For the EN-CS task, the Recall improvements of Biography questions over Definition questions, Definition questions over Relation question, and Relation questions over Event questions are about 17.0%, 5.0%, and 24.0%, respectively. The Recall improvements in the CS-CS task are also significant.
- Table 3 also indicates that the RUN-1 in the CS-CS cross-lingual task outperforms that in the EN-CS monolingual task. However, the difference is not significant. The reason might be: there are few translation errors in the EN-CS task. For example, 60% translations of questions use the same words as that in the monolingual counterparts; 30% of the translations are correct even though they use different words from their monolingual counterparts; only 10% translations include at most an error.

[2] The best system is from CSWHU who does not participate the EN-CS task.

About the RUN-2 and the RUN-3, Table 4 summarizes the F-score performances based on human evaluation.

Table 4. The F-scores (%) of the RUN-2 and the RUN-3

	RUN-2		RUN-3	
	EN-CS	CS-CS	EN-CS	CS-CS
Event	14.08	14.07	8.09	10.86
Definition	23.37	25.65	12.57	16.18
Biography	30.27	32.53	20.77	18.06
Relation	22.80	23.76	12.10	16.50
all	21.79	22.98	12.73	15.06

Comparing Table 4 with Table 3, we can make the following conclusions.

- The SVM-based models, i.e., the RUN-1 and the RUN-2, are significantly superior to the centroid-vector model, i.e., the RUN-3. For the EN-CS task, the F-score improvements of the RUN-1 and the RUN-2 over the RUN-3 are about 9.0%. For the CS-CS task, both the RUN-1 and the RUN-2 improve the F-score of the RUN-3 by about 8.0%.
- As mentioned above, the RUN-2 uses the five external resources to construct target profiles, while the RUN-1 uses only one single external resource. Kor, et al. [4] showed that target profiles constructed from multiple external resources outperform that constructed from one single external resource in the centroid-vector model. However, Table 4 and Table 5 demonstrate that the impact of target profiles are not of significance in the case of the SVM-based model. As a result, there is no need to mine external resources as many as possible for the SVM-based model.

The CCLQA track also provides automatic evaluations as a supplementary of human evaluation. The F-scores of the three runs are shown in Table 5.

Table 5. The F-scores (%) using automatic evaluation

	RUN-1	RUN-2	RUN-3
the EN-CS task			
Event	19.52	19.12	18.01
Definition	20.95	22.99	19.14
Biography	35.81	36.40	34.32
Relation	32.18	32.07	28.19
all	26.86	27.24	24.55
the CS-CS task			
Event	17.89	18.49	16.42
Definition	24.24	26.14	19.86
Biography	37.22	36.86	32.75
Relation	34.00	33.84	31.10
all	27.85	28.30	24.78

The automatic evaluation results in Table 5 indicate that:

- The F-scores of the RUN-2 in terms of the EN-CS task is 27.24%, the best system among all participants' systems. The next best system achieves 22.90% F-score. For the CS-CS task, the RUN-2 achieves 28.30%, ranked in second place. The F-score of the best system is 37.75%.
- The ranking of the runs in both the EN-CS task and the CS-CS task is consistent, i.e., RUN-2 > RUN-1 > RUN-3. Even though the RUN-2 consistently outperforms the RUN-1 in the EN-CS task and the CS-CS task, the differences are not significant. For example, the improvements of the RUN-2 over the RUN-1 in the EN-CS task and the CS-CS task are 0.38%, and 0.45%, respectively. Therefore, the conclusion from this table is same as that from Table 4: target profile does not significantly impact the performances of the SVM-based model, which plays an important role in the centroid-vector model [4].
- According to this table, the difficulty ranking of answering complex questions is: Event > Definition > Relation > Biography. It is a little bit strange that this ranking is different from that in Table 3.
- The F-scores of the CS-CS task outperform that of the EN-CS task by around 1.0%, which is different from human evaluation. Although the automatic evaluation cannot conduct concept-matching between system responses and standard answers, and is just a supplementary of the human evaluation in the CCLQA track, it has good consistency among systems, which is hard to obtain for human evaluation.

6.2 IR4QA + CCLQA Track

As Figure 1 illustrated, the *document-retrieval* module may impact the performance of the CCLQA system. To find out the best IR strategy that brings about the best end-to-end QA performance, the CCLQA track required us to submit our systems based on the retrieval results provided by the participants' systems of the IR4QA track. We submitted six EN-CS cross-lingual runs that based on the retrieval results named CMUJAV-EN-CS-01-T-limit50 (CJAV1), CMUJAV-EN-CS-02-T-limit50 (CJAV2), and MITEL-EN-CS-01-T-limit50 (MITEL). The IR4QA results of the CJAV1, CJAV2, and MITEL [7] in terms of the test question set are shown in Table 6.

Table 6. The IR4QA results of the CJAV1, CJAV2, and MITEL

	Mean AP	Mean Q	Mean nDCG
MITEL	36.82	35.18	52.06
CJAV1	34.52	32.58	47.89
CJAV2	33.87	31.99	47.25

Using them, the automatic evaluation F-scores of our runs are shown in Table 7.

Table 7. The F-scores of our runs based on the CJAV1, CJAV2, and MITEL

	CJAV1	CJAV2	MITEL
RUN-1	27.63	27.30	27.50
RUN-3	24.91	24.44	24.47

Comparing Table 6 and Table 7, we find that:

- As Table 7 indicates, the performances of the CJAV1 is better than that of the CJAV2, thus, the RUN-1 and the RUN-3 based on the CJAV1 outperform that based on the CJAV2, as Table 8 indicates.
- Even though the MITEL is better than the CJAV1, the CCLQA runs based on the MITEL is slightly worse

than the runs based on the CJAV1. This is out of our expectation. We try to find out the reason in the following section.

To understand the possible reason, Table 8 summarizes the performances of the MITEL, the CJAV1, and the CJAV2 over types of questions. Table 9 breaks the performances of the RUN-1 based on the MITEL, the CJAV1, and the CJAV2 in accordance with types of questions. The less-than sign (<) denotes that the left performance is lower than the right, and the greater-than sign (>) indicates the left performance is better than the right.

Table 8. The Mean-AP scores of the CJAV1, the JCAV2 and the MITEL over types of questions.

	MITEL		CJAV2		CJAV1		MITEL
Event	26.57	>	19.47	<	19.53	<	26.57
Definition	37.10	<	49.85	>	48.65	>	37.10
Biography	45.15	<	46.15	<	46.65	>	45.15
Relation	41.37	>	29.43	<	32.00	<	41.37

Table 9. The performances of the RUN-1 based on the CJAV1, the CJAV2, and the MITEL over types of questions.

	MITEL		CJAV2		CJAV1		MITEL
Event	18.92	>	**18.74**	**>**	**17.74**	<	18.92
Definition	17.23	<	**22.46**	**<**	**23.08**	>	17.23
Biography	**37.87**	**>**	**36.17**	<	38.37	>	37.87
Relation	36.01	>	33.18	<	33.40	<	36.01

From the tables, we can conclude that:

- The performances of the IR4QA systems basically have consistent influences over the types of questions in the RUN-1. For instance, the CJAV1 outperforms the MITEL in terms of Definition and Biography questions. Therefore, the performances of Definition and Biography questions of the RUN-1 based on the CJAV1 are better than that based on the MITEL. There are three exceptions in Table 9 that are marked out with bold fonts.
- However, the extent of the impact of the IR4QA system on different types of questions are not the same in the RUN-1. For instance, the Mean-AP improvements of the MITEL over the CJAV1 in terms of Event and Relation questions are about 7%, and 9%, respectively, however, the RUN-1 based on the MITEL improves that based on the CJAV1 just by 1.2%, and 2.6%. This results in the phenomenon in Table 7: the performances of the RUN-1 based on the MITEL is lower than that based on the CJAV1.

7 Summary

We implemented two models, i.e., the SVM-based model and the centroid-vector model, to answer complex questions in the NTCIR-7 CCLQA track. The centroid-vector model aims to exploit external resources as many as possible to construct accurate target profiles, and then calculate similarities between candidate answers and target profiles using vector space model. The SVM-based model, however, focuses on learning evidences of candidate answers (as described in Section 5.2), and then uses an SVM classification model to incorporate the learned evidences to improve complex question answering.

Our results for complex cross-lingual questions are satisfactory compared with the results of other groups. However, much work remains, there is still much space for improvement, especially for Event question.

References

[1] C. C. Chang and C. J. Lin. Libsvm: a library for support vector machines. *http://www.csie.ntu.edu.tw/ cjlin/*.

[2] Y. Chen, M. Zhou, and S. Wang. Re-ranking answers for definitional qa using language modeling. *Proc. of ACL/COLING-2006*, 2006.

[3] H. Cui, M. Y. Kan, and T. S. Chua. Unsupervised learning of soft patterns for definition question answering. *Proc. of WWW2004*, 2004.

[4] K. W. Kor and T. S. Chua. Interesting nuggets and their impact on definitional question answering. *Proc. of SIGIR-2007*, pages 335–342, 2007.

[5] T. Mitamura, E. Nyberg, H. Shima, and et al. Overview of the ntcir-7 aclia: Advanced cross-lingual information access. *Proc. of NTCIR-2008*, 2008.

[6] M. Pasca. Answering definition questions via temporally-anchored text snippets. *Proc. of IJCNLP2008*, 2008.

[7] T. Sakai, N. Kando, C. J. Lin, T. Mitamura, and et al. Overview of the ntcir-7 aclia ir4qa subtask. *Proc. of NTCIR-2008*, 2008.

[8] Y. Z. Wu, J. Zhao, X. Bo, and H. Yu. Chinese named entity recognition model based on multiple features. *Proc. of HLT/EMNLP-2005*, pages 427–434, 2005.

[9] J. X. Xu, A. Licuanan, and R. Weischedel. Trec2003 qa at bbn: Answering definitional questions. *Proc. of TREC2003*, 2003.