Complex Cross-lingual Question Answering as a Sequential Classification and Multi-Document Summarization Task

Hideki Shima, Ni Lao, Eric Nyberg, and Teruko Mitamura
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Abstract

In this paper, we describe the JAVELIN IV system, which treats complex question answering as a sequential classification and multi-document summarization task. Our research and development effort is based on various forms of linguistic annotation, and a comparison of various answer extraction and summarization algorithms. We discuss the use of different units of extraction, the effect of different syntactic features for classification, and the effect of different summarization strategies. We also analyze how the performance of machine translation and information retrieval affect the performance of question answering. In the NTCIR-7 CCLQA main track official evaluation, our system achieved 16.3% and 19.2% accuracy in the English-to-Japanese and English-to-Chinese subtasks, respectively.

1. Introduction

Previous QA research at CMU focused on monolingual English factoid QA (JAVELIN I), monolingual English complex QA (JAVELIN II) and cross-lingual factoid QA (JAVELIN III)[1][30]. All JAVELIN configurations use the same modular architecture (see Section 2). The system we describe here, JAVELIN IV, conforms to the same architecture but implements a novel approach that is distinct from JAVELIN III, which was developed for factoid QA[1]).

The rest of this paper is structured as follows. We present the JAVELIN IV architecture and modules in Sections 2 through 4. Then Sections 5 and 6 present our results from the formal evaluation, and discuss the interesting issues we discovered. We conclude in Section 7 with some ideas for future research.

2. Javelin Architecture

JAVELIN's modular architecture consists of four modules: 1) the Question Analyzer (QA), which is responsible for analyzing question inputs to determine the information need; 2) the Retrieval Strategist (RS), which is responsible for retrieving a ranked list of possible answer-bearing documents; 3) the Information eXtractor (IX), which is responsible for extracting and scoring/ranking answer candidates; and 4) the Answer Generator (AG), which is responsible for final answer generation and provides duplicate answer removal (see Figure 1).

The JAVELIN architecture has two notable characteristics. First is the language-independent design. All the algorithms in JAVELIN IV are designed in a language-independent way, using uniform module interfaces for the machine translation (MT) and other NLP modules used to process question and answer texts. Each module loads language-specific resources such as dictionaries, question and answer patterns, and trained classifiers. The second characteristic of our architecture is the use of distributed computing. We deploy machine translation, passage retrieval, and text processing (e.g. parsing) as distributed services, using techniques including SOAP, Java RMI, and TCP/IP sockets.

This paper will focus on the information extraction and answer generation components of JAVELIN IV. Further details regarding the question analysis and retrieval modules can be found in our NTCIR-7 IR4QA paper [2].

2.1. Experimental Settings for System Development

During design and implementation of the system, we conducted a variety of experiments on the ACLIA training dataset, which contains 4 types of topics: DEFINITION, BIOGRAPHY, RELATIONSHIP and EVENT. These experiments are described along with the description of each relevant module, in Section 3. We used the nugget pyramid F3 metric (with automatic nugget evaluation) for preliminary evaluation and tuning of the IX and AG modules.

For Japanese we use the ACLIA training set (101 topics) for training purposes. Prior to generating our formal run results, we used all the training topics to train the classifiers for answer extraction (Section 3). We used POURPRE [3] for automatic nugget evaluation; nuggets receive a score of 0 or 1 if the POURPRE score is above a threshold value of 0.5.

For Chinese we use the ACLIA training set (88 topics) for training purposes. For each type of question, 7 topics are randomly picked for testing and are held out; the remaining topics are used for training. During automatic evaluation, a nugget is considered matched if one of its clauses is matched in the system output text. Clauses are created by splitting nuggets at comma boundaries.

[1] A simple extension of JAVELIN III was evaluated on monolingual Japanese complex QA in the NTCIR-6 QAC track [10], but not continued; JAVELIN IV is a completely new implementation.

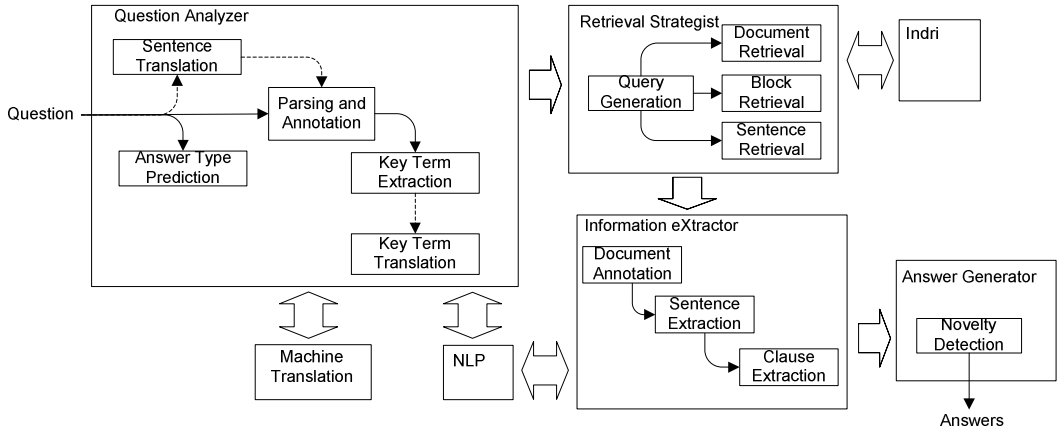

Figure 1 JAVELIN IV Modular Architecture

3. IX Module: Answer Extraction as a Sequential Classification Task

The IX module is responsible for extracting answer candidates given texts retrieved by the RS, and assigning scores to each candidate so that a ranked list of candidates may be produced as output. To the best of our knowledge, this work is the first to treat answer extraction for complex QA as a sequential classification problem.

Traditionally, state-of-the-art complex QA (mostly definitional QA) systems have adopted a sentence retrieval/extraction approach [4][5][6][7][23]. The same approach was taken in our first complex QA system (for monolingual Japanese) which was entered in the NTCIR-6 QAC-4 track [9]. According to the TREC 2003 Definitional QA task results, a simple sentence-based baseline was good enough to outperform all but one of the other 15 runs in the track.

However, answers to complex questions are often longer than a single sentence; answers can span multiple sentences, and a per-sentence answer analysis may not yield the best results. A study by Verberne [10] created 1177 *why*-questions from the web, and randomly picked 400 of them in order to analyze the distribution of answer-bearing text units. From Wikipedia, answers to 54% of the questions were found, and only 13% of these answers were one sentence long; 81% are longer than one sentence and shorter than one paragraph. Although the *why*-questions studied by Verberne are different from the four types of questions examine here, evidence suggests that a bag-of-sentences view of a document is not appropriate for complex QA.

Table 1 Answer-bearing document JA-010104167 for ACLIA1-JA-D362 "Who is Kitaoji Rosanjin?"

Line	Original Japanese document (excerpt)
1	◇**北大路魯山人**（きたおおじ・ろさんじん）
2	１８８３年京都市生まれ。
3	陶芸家。
4	東京・赤坂に会員制の高級料亭「星岡茶寮」を開設し、顧問兼料理長として美食家の名をはせる。

Line	Translation of above excerpt
1	◇**Rosanjin Kitaoji** (*kana readings for Rosanjin Kitaoji*)
2	Born in Kyoto in 1883.
3	Ceramist.
4	Opened the Hoshigoaka Restaurant in Tokyo, served as the master chef and gained a reputation as a gourmet.

We also found cases where sentence-based analysis is problematic for our training topics and corpus. In Table 1, an excerpt from an answer-bearing document is shown for a biography question, with key terms in bold face. A sentence containing the key term is merely the title of an expositional paragraph, and the key term never matches the answer-bearing sentences; see lines 2-4 in Table 1.

We hypothesize that modeling the answer context using an inter-sentential dependency analysis solves this problem; we test this hypothesis using a sentence-level sequential learning method. Our approach uses supervised classifiers and classifier learners implemented in the MinorThird package [11]. Two sequential frameworks we adopt are CMMs, a type of directed graphical model known as Conditional Markov Models [18]; and an undirected graphical model known as Conditional Random Fields (CRFs). In CMM model (Figure 2), a sequence of binary states $\mathbf{s} = \{s_1,...,s_n\}$ (i.e. answer-bearing or not) and a given observation sequence $\mathbf{o} = \{o_1,...,o_n\}$ consisting of overlapping features are the random variables, and a base learner (described later) is extended by incorporating previously predicted classes as features[2]. At training time, modeling effort is spent on estimating $p(\mathbf{o}|\mathbf{s})$ in a discriminative framework like CMMs. At classification time, $\mathbf{s}$ is found by maximizing the prediction confidence given $\mathbf{o}$.

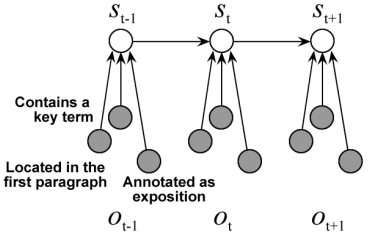

Figure 2 Graphical model for CMMs.

At training time, models are trained for each answer type using monolingual inputs. The gold standard labels are assigned from answer-bearing passages in the ACLIA training dataset, extracted by topic developers. Regarding the input to the IX module, we mix documents retrieved from the RS module with the answer-bearing documents

[2] History size = 3

to control the positive-negative example ratio and also to utilize the full potential of the training resource.

At test time, the IX produces a ranked list of answer candidates, i.e. classified instances with a positive label ranked according to the classification scores.

3.1. Analysis on Japanese Training Data

Using the base learners listed below, we compared non-sequential (hereafter local) and sequential models.

- **CRFs**[3] are linear-chain sequential models based on the CRF algorithm [12][13].
- **Passive Aggressive (PA)**[4] [14] is originally an online algorithm, which we used in batch mode.
- **SVM** [5] is a margin-based classifier learning algorithm[15][16].
- **Voted Perceptron (VP)**[6] [17] is similar to SVM in the sense that it is based on large-margin linear classifier.

In addition, we compared normal and "tweaked" versions of CRFs, PA and VP. We tweaked training for PA and VP by adjusting the learner's parameters to favor more recall, and optimizing the score according to the F-measure given beta[7]. Tweaked training for CRFs is done by adjusting the bias term of the hyperplane[8].

Table 2 Classifier Learner Comparison measured by automatic metric. NOR=normal training, TWE=tweaked training, LOC=local model, SEQ=sequential model.

	BIO	DEF	EVE	REL	ALL
Baseline	0.66	0.42	0.44	0.51	0.51

		BIO		DEF		EVE		REL		ALL	
		LOC	SEQ	LOC	SEQ	LOC	SEQ	LOC	SEQ	LOC	SEQ
CRFs	NOR	-	0.54	-	0.42	-	0.24	-	0.45	-	0.43
	TWE	-	0.65	-	**0.52**	-	0.45	-	0.56	-	0.57
PA	NOR	0.65	0.55	0.44	0.39	0.16	0.05	0.51	0.39	0.48	0.38
	TWE	0.69	0.70	0.49	0.49	0.38	0.41	0.65	**0.79**	**0.59**	**0.59**
SVM	NOR	0.59	0.62	0.44	0.30	**0.59**	0.21	0.49	0.58	0.49	0.45
VP	NOR	0.60	0.43	0.42	0.33	0.23	0.03	0.53	0.23	0.47	0.29
	TWE	0.67	**0.71**	0.46	0.48	0.45	0.43	0.68	0.64	0.57	**0.59**

To produce the experimental results shown in Table 2, we compared automatic metric scores on IX output (top 15 answer candidates per topic, generated from multiple classifiers in 5-fold cross validation) and the baseline algorithm. The baseline is a simple sentence-based algorithm, like the SENT-BASE algorithm presented at TREC 2003 or the Organizer run algorithm [8]. Numbers in bold face are the best results for a particular answer type.

[3] History size = 1
[4] Used in classification mode. Insensitivity parameter eta=1, unrealizable case parameter gamma=0.1, and voting scheme is enabled
[5] linear-kernel C_SVC SVM mode with a default parameter set in libsvm
[6] Epoch size = 5
[7] Beta = 3 for favoring more recall
[8] Bias = -1.2 for favoring more recall

Contrary to the results from TREC 2003, where the baseline outperformed most of the systems, here the best scores under each answer type are higher than the baseline. Given this evidence, we found models were best tuned when trained in tweaked and sequential mode. Especially, we found that tweaked sequential VP was consistently good on the sample questions, and decided to use it in formal run.

The features we used in the above experiment and in the formal run are listed below:

- **Key term**: fires when at least one key term observed.
- **Context**: fires for terms found around the key term.
- **Enclosure**: fires when a key term is observed and it is enclosed with quotations for an emphasis.
- **Position**: offset of the sentence within a paragraph.
- **Alias**: fires when an ALIAS pattern is applicable. Patterns are lexico-syntactic patterns learned using a bootstrapping method (see details in [2]).
- **Cue**: fires when an answer type specific hand-crafted cue exists. Among cues, useful ones are categorized as strong cues (e.g. 誕生 (*was born in*), 卒業 (*graduated from*)), less co-occurring ones are categorized as weak cues (e.g. 死去(*died*), 受賞 (*awarded)*), and negative indication of answers are categorized as negative cues (e.g. ■写真説明(*photo explanation*)).
- **Coreference**: fires when a coreference annotation exists.
- **Exposition**: fires when an exposition annotation exists. Sentences in Table 1 are good examples where exposition annotations are applicable.

Features were designed based on observations over answer-bearing documents. For example, answers to DEF and BIO questions often contained both key term and its alternative form (i.e. alias), which motivated us to implement the pattern based alias feature extractor. As another example, intuitively, coreference resolution is important in analyzing inter-sentential dependencies. In Japanese, once a referent is explicitly shown, subsequent sentences often omit a reference to it. Given that zero-pronoun phenomena, we annotated a sentence if Subject is missing but earlier in the same paragraph another sentence exists such that the key term is Subject.

In addition, we also generated history/future features which are copied from other sentences in the same paragraph.

3.2. Analysis on Chinese Training Data

For Chinese, we decided to use a maximum entropy (Maxent) model trained in CMMs, a.k.a. Maximum Entropy Markov Models [19], based on results from a preliminary experiment like those shown in Table 2.

The features extracted from each sentence are:

- **Key terms**: number of times matched
- **Named entity**
- **Cue terms**

- **Language phenomena** annotated by hand-written rules:
  - **Subject:** the NP that directly attaches to and precedes the head word of the sentence.
  - **Anaphora:** annotate sentences (clauses) that have no subject with the subject of the previous sentence (clause).
  - **Co-reference:** any pronoun word like 她(*she*), 他(*he*) 他们(*they*), etc is annotated with the subject of the previous sentence (clause).
  - **Possessive:** find the pattern "NP 的 NP"
  - **Apposition:** find the pattern "NP, NP"
  - **Be verb:** find the pattern "NP 是"
- **Co-occurrences** of Named entity and cue terms

Examples of some of the model weights are shown in Table 3. We can see that anaphora is not welcomed in biography and definition questions, as their answers usually have the form "XXX is …" "XXX does …", etc. which do not contain anaphora. Anaphora within a sentence is a positive indicator for event and relation questions.

Being the first clause in a sentence is a positive indication for biography and definition questions.

Death related words like 死于(die at),死(die), 亡(die), 去世(pass away), 逝世(pass away), 过世(pass away) etc. are a positive indication for biography questions. They are also a positive indication for event and relational questions (but only when sentences are the extraction unit).

Table 3 Model weights for some features

Setting		DEF	BIO	REL	EVE
		ANAPH: anaphora			
clause	CC	-0.60	-2.00	0.19	0.39
	EC	-0.90	-2.30	0.48	0.25
sentence	CC	-0.50	-1.00	0.00	0.00
	EC	-0.10	-0.30	0.00	0.00
		SentFirst: being the first sentence/clause in a paragraph			
clause	CC	0.37	0.19	-0.30	-0.00
	EC	0.54	0.19	-0.30	0.01
		DEATH: presence of death words			
clause	CC	-1.70	0.97	0.00	-0.00
	EC	-0.70	0.78	0.00	-1.10
sentence	CC	-0.50	0.44	0.57	0.49
	EC	-0.20	0.39	0.05	0.17
		PAST: presence of past tense words			
clause	CC	-0.10	0.58	-1.30	0.16
	EC	-0.60	0.57	-1.90	0.22
sentence	CC	-0.50	0.30	-1.00	-0.30
	EC	-0.40	0.23	-1.00	-0.10

Words indicating past tense like 已经(*already*), 曾经(*had*), 早已(*already*), 早就(*already*), 已(*already*), 曾(*had*), 就(*than*) 一度(*once*), 现已(*already*) etc. are a positive indication for biography questions. They are a positive indication for event questions only when clauses are the extraction unit.

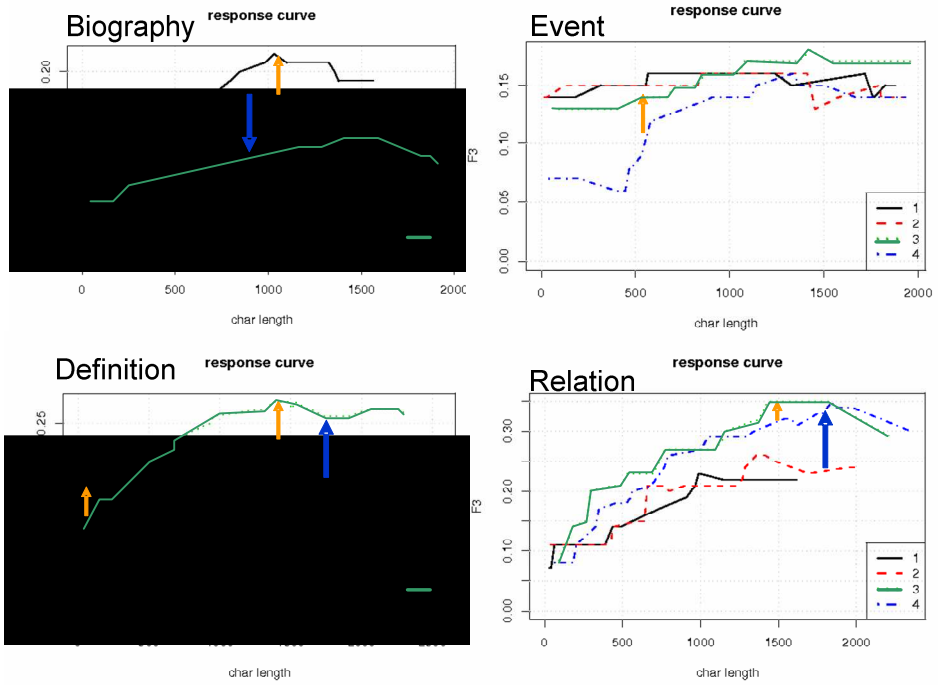

Figure 3 Answer Extraction Strategies. Horizontal axis is the number of characters of top ranked system responses. Vertical axis is the IX module evaluation (F3). Curves are 1) clause, deep 2) clause, shallow 3) sentence, deep 4) sentence, shallow.

From experiments on training data (Figure 3), we can see that deep sentence analysis always outperforms shallow analysis. Furthermore, the sentence seems to be a better extraction unit for more complex questions (Event and Relation) but not the simpler ones (Biography, Definition), where the clause is a better extraction unit.

However, a clause might not have enough information for human to evaluate it in isolation (e.g. when anaphora and/or co-reference occurs). This hypothesis could be tested via human evaluation.

4. AG Module: Answer Selection as a Text Summarization Task

After answer candidates are extracted by the IX, our interest is in removing duplicate information from the candidate list, in order to avoid being penalized for returning long, redundant responses. The Answer Generation (AG) module is responsible for this selection task.

Unlike factoid QA, where word-level answers are easy to de-duplicate, the research question in complex QA task is how to detect the duplications among longer answers, which are sentence-level or sometimes even passage-level. Ranking answer passages is analogous to selecting vital passages for text summarization. This leads to the hypothesis that techniques from text summarization can be adapted to the Answer Selection problem CLQA. Developing this into an operational hypothesis, we posit that proposed approach A) is better than the baseline B):

A) Automatically assign scores to the AG output as a result of sentence-level Multi-Document Summarization (MDS) process from the IX output into *N* answer candidates
B) Automatically evaluate score on top *N* answer candidates in the IX output

According to an experiment on training data, we set *N* to be 15 for JA and 50 for CS in an attempt to maximize the final score.

As a method, we introduced Maximal Marginal Relevance (MMR) [20][21] well known as query-focused MDS algorithm in Text Summarization. The algorithm has an advantage that it is simple, general-purpose and language-independent.

Formally, let *D* be a document (i.e. list of summary candidates), *Q* be a query, *S* be a summary (i.e. list of candidates selected as summary) and *r* be a parameter to balance how much to balance duplication effect. MMR iteratively selects best summary candidate as shown in (1) where a summary candidate is selected such that maximizes a similarity to Q subtracted by a similarity to current summary. The selected candidate is added to the summary and iteration continues until convergence when summary grows to a specified size.

$$MMR = \arg\max_{D_i \in R \setminus S} \left[\mathrm{Sim}_1(D_i, Q) - r \max_{D_j \in S} \mathrm{Sim}_2(D_i, D_j) \right] \quad (1)$$

For the answer generation task, we view a summary candidate as an answer candidate from the IX, the first similarity score as the IX score for the answer candidate.

For calculating the second similarity term, we segmented answer candidates into characters and calculated the position-independent error rate (PER) among them. PER is an automatic evaluation metric used in Machine Translation evaluation. It is similar to the word error rate but instead uses a position independent Levenshtein distance (bag-of-word based distance) [22].

Based on the changes above, our customization can be formalized as shown in (2).

$$MMR = \arg\max_{A_i \in R \setminus S} \left[\mathrm{IX}(A_i, Q) - r \max_{A_j \in S} (1 - \mathrm{PER}(A_i, A_j)) \right] \quad (2)$$

We compared an implementation to the baseline on the JA training dataset, changing the *r* parameter. As the results show in Figure 4, MMR outperforms the baseline, when an appropriate *r* is found. In this experiment, we found *r*=0.6 maximizes the score, and thus we decided to use that *r* value in the formal run, hoping the hypothesis still holds.

There are some existing works (e.g. [23][24][25]) where Text Summarization techniques are used for QA.; it is also true that a Question Answering system has been used for Text Summarization [26].

Our work is unique in the sense that we used a Text Summarization technique exclusively for the AG module, and not for the entire QA system. In this way, we can easily see the pure effect of summarization algorithm as we saw in Figure 4, or we can even implement and switch to another answer selection mode such as a probabilistic model [27][28].

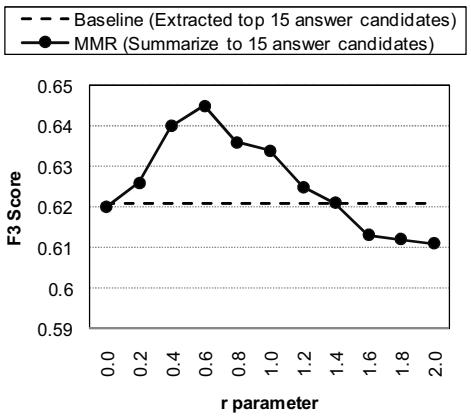

Figure 4 MMR-based AG vs baseline.

5. CCLQA Main Track Result and Analysis

5.1. Analysis on Japanese Results

The answer type classification accuracies were 60% and 73% for EN-JA and JA-JA respectively. The confusion matrix in Table 5 shows there are particular kinds of errors, especially salient cases include: 24 EVENT types were predicted as RELATIONSHIP in EN-JA, and 18 EVENT were predicted as DEFINITION in JA-JA. In both cases, questions were asked in unseen way as in training dataset. For instance, an EVENT question from Topic 151 is シドニー五輪の聖火リレーについて教えてください。(*Please tell me about the Sydney Olympic Torch Relay*).

Table 4 Human-in-the-loop scores for the Javelin runs in CCLQA Main Track.

EN-CS runs	DEF	BIO	REL	EVE	ALL
CMUJAV-EN-CS-01-T	0.2129	0.2678	0.1864	0.1346	0.1924
CMUJAV-EN-CS-02-T	0.1309	0.1259	0.1032	0.0662	0.1022
CMUJAV-EN-CS-03-T	0.2192	0.2324	0.2145	0.1345	0.1950

CS-CS runs	DEF	BIO	REL	EVE	ALL
CMUJAV-CS-CS-01-T	0.2326	0.2498	0.2301	0.1219	0.2021
CMUJAV-CS-CS-02-T	0.1255	0.0897	0.1330	0.0739	0.1051
CMUJAV-CS-CS-03-T	0.2305	0.2066	0.2682	0.1527	0.2137

EN-JA runs	DEF	BIO	REL	EVE	ALL
CMUJAV-EN-JA-01-T	0.3772	0.1250	0.1641	0.0433	0.1627
CMUJAV-EN-JA-02-T	0.3701	0.1388	0.1667	0.0510	0.1671
CMUJAV-EN-JA-03-T	0.3712	0.1083	0.1210	0.0395	0.1440

JA-JA Runs	DEF	BIO	REL	EVE	ALL
CMUJAV-JA-JA-01-T	0.3980	0.1749	0.2291	0.0813	0.2077
CMUJAV-JA-JA-02-T	0.3918	0.1843	0.2205	0.0712	0.2027
CMUJAV-JA-JA-03-T	0.3935	0.1774	0.1987	0.0728	0.1956

Table 5 Confusion matrix for answer type classification results from JA formal runs.

EN-JA		Predicted			
		DEF	BIO	REL	EVE
Actual	DEF	18	0	2	0
	BIO	1	11	8	0
	REL	4	0	26	0
	EVE	1	0	24	5

JA-JA		Predicted			
		DEF	BIO	REL	EVE
Actual	DEF	20	0	0	0
	BIO	2	18	0	0
	REL	4	1	25	0
	EVE	18	0	2	10

Table 6 Top 5 system responses of JA-JA-01 run for the question "What is the Kyoto Protocol?" (ACLIA1-JA-T25).

Rank	System responses
1	**京都議定書**は国際交渉の画期的成果と考えられた。
2	**京都議定書**は先進国全体で２０１０年までに９０年比で温室効果ガス排出量の５%削減を義務付け、国別の削減目標も明記した。
3	温暖化防止の基本になるのは、１９９７年に開かれた気候変動枠組み条約第３回締約国会議（ＣＯＰ３、地球温暖化防止京都会議）で採択された**京都議定書**だった。
4	０８～１２年の間にＣＯ２など温室効果ガス排出量を先進国全体で９０年より５・２%減らすと決め、ＥＵ全体で８%、米国で７%、日本で６%など国ごとの削減目標値も定めた。
5	だが途上国に義務を課さなかったことに米国は反発し、今年３月に不支持を表明した。

Rank	System responses (translated)
1	The **Kyoto Protocol** has been thought to be a revolutionaly outcome of international negotiation
2	The **Kyoto Protocol** set obligations for industrialized countries to reduce their collective GHG emissions by 5% during 1990 and 2010, and specified target reduction rate for individual country.
3	Basis of global warming prevention program would be the **Kyoto Protocol**, adopted at COP3
4	Decided to reduce emission of industrialized nation's green house gas (for example CO2), by 5.2% during 2008 to 2012 as compared to 1990, and defined goal values, such as EU for 8%, US for 7% and Japan for 6%.
5	However, in March US showed a disapproval for not setting obligations to developing countries.

In the main track, we submitted runs from three different learning frameworks (see details in 3.1).

- **CMUJAV-EN-JA-01, CMUJAV-JA-JA-01**: classify answer-bearing sentences with tweaked (recall-optimized) sequential model.
- **CMUJAV-EN-JA-02, CMUJAV-JA-JA-02**: classify answer-bearing sentences with tweaked non-sequential model.
- **CMUJAV-EN-JA-03, CMUJAV-JA-JA-03**: classify answer-bearing sentences with normal local (non-sequential) model.

Comparing the results in Table 4, we can see scores for the 01 and 02 runs are higher than 03 run, indicating that inter-sentential analysis by sequential classification contributes to a complex QA system. Table 6 shows the actual system responses where you can see benefits from our sequential analysis; system responses ranked 4th and 5th are virtually impossible to return for traditional QA systems as they do not contain any matching key terms.

5.2. Analysis of Chinese Results

Our answer type classification accuracy was 87% and 88% for EN-CS and CS-CS respectively [8]. From the confusion matrix in Table 7, we learn that most error comes from classifying event questions as other types. One reason is that our manually-crafted patterns and weights for question classification are still not robust enough. Another reason is the inconsistent definition of answer type as used by the training data set and testing data set. For example, the following questions have very similar structure, but are classified differently in training and testing data.

Relation question in training data:

D73 战争对于世界石油供应有何影响？(*What is the relationship between wars and the world oil supply?*)

D83 美国通过对华永久正常贸易关系法案，对中国有何重大关系 (*How important to China is America's passage of the permanent normal trade relations bill?*)

Event question in testing data:

T44 举出 911 事件对美国的影响。(*List the impact of the 911 incident on the United States.*)

T46 列举亚洲金融危机对经济的影响。(*List the impact of the Asian financial crisis on the economy.*)

Table 7 Confusion matrix for CC formal run answer type classification result.

EN-CS		Predicted			
		DEF	BIO	REL	EVE
Actual	DEF	20	0	0	0
	BIO	0	20	0	0
	REL	0	0	27	3
	EVE	5	0	5	20

CS-CS		Predicted			
		DEF	BIO	REL	EVE
Actual	DEF	20	0	0	0
	BIO	0	20	0	0
	REL	0	0	30	0
	EVE	3	1	8	18

For Chinese runs we compared strategies to explore the proper extraction unit for the complex QA task: a) **CMUJAV-EN-CS-01, CMUJAV-CS-CS-01**: extract clauses; b) **CMUJAV-EN-CS-02, CMUJAV-CS-CS-02**: extract sentences; c) **CMUJAV-EN-CS-03, CMUJAV-CS-CS-03**: extract sentences and clauses by breaking a sentence into clauses, and then running the sequential model on them. The clauses selected by the model are concatenated together to form the output.

The results of these runs are shown in Table 4. First we can see that using sentence as the extraction unit results in very low performance. This was to be expected, because many sentences are much longer than a clause, and such cases hurt our precision score. However, what's not expected is that sentence extraction has not only lower precision but also lower recall. We found several cases in the formal run result where although the same clause was labeled relevant by the human evaluator in clause-extraction output, but not in sentence-extraction output. Therefore, we assume that, since clauses are shorter, human evaluators are less likely to miss their matched nuggets. Another possible reason is that when a sentence is longer, it's harder for the designed features to capture its relevance.

Second, we can see that extracting clauses from sentences is slightly better than directly extracting clauses (Table 4). From a closer look at the results we know that sentence-clause extraction has slightly lower precision, but higher recall. Note that sentence-clause extraction performs significantly better on more complex questions like relation and event, but clause extraction is better for biographic questions. This indicates that we can improve overall performance by switching strategies according to question type.

Although human assessors have no problem understanding clauses as the extraction unit, we should be cautious about concluding that clause is a better extraction unit than sentence, because real QA users tend to require contextual background for the system responses [29].

After a close look into the results we found that the feature "first clause of a sentence" helps clause-extraction approach beat sentence-clause extraction approach in many biography questions. Now we know a better design would be to train two models for the sentence-clause extraction strategy, one to extract sentence, and another to extract clauses from sentence which are marked with the feature "first clause" (of a sentence).

Third, from the scatter plots (Figure 5) we can also see that in quite a few topics clause extraction or sentence extraction received a zero score while sentence-clause extraction did not. This indicates the robustness of sentence-clause extraction.

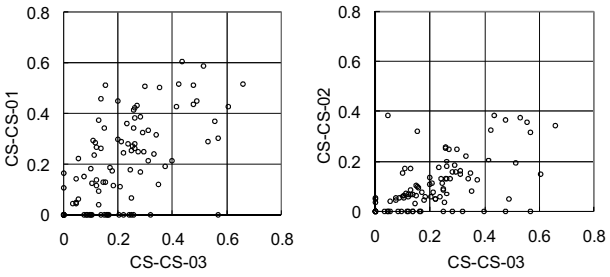

Figure 5 Official score plots for sentence-clause extraction (horizontal) verses clause extraction (left) and sentence extraction (right).

Now we analyze the effect of machine translation and the correlation between the choice of IR component and the quality of the final QA output (Table 8). The topics are also categorized by how IR (our IR4QA run 02 [2]) and QA (our CLQA run03) scores are affected by translation; this implies that they improve (+), remain the same (=), or decrease (-).

Let's first consider biography and definition questions. In most cases, translation has almost no affect on IR scores, and some affect on QA scores. Since the questions are simple, both monolingual and cross lingual systems are very likely to correctly identify the key term, therefore IR scores don't change much. However, the extra key terms that translation brings in can affect answer extraction in both positive (for biography questions) and negative (for definition questions) ways.

Consider the event and relation questions. Since the questions are longer, both monolingual and cross lingual systems are likely to experience errors in identifying the key term; therefore IR scores can change significantly in positive or negative ways. Note that even for the cases with positive change is IR scores, the QA score is mostly dropping. One reason for this contradiction is that even though translation brings in extra relevant documents that were not seen in the data creation process, the nuggets in these documents are no added into the gold standard nugget pool. On the other hand, a greatly reduced IR score does not necessarily lead to a worse QA score. This also indicates the incompleteness of IR annotation.

Table 8 Translation's effect on IR and QA. "IR =" represent the cases with MAP score change smaller than 1%, "QA =" the cases with F3 score change smaller than 1%.

BIO	QA -	QA =	QA +	EVENT	QA -	QA =	QA +
IR -	0	0	1	IR -	5	0	7
IR =	3	5	11	IR =	2	2	1
IR +	0	0	0	IR +	6	2	5
DEF	**QA -**	**QA =**	**QA +**	**REL**	**QA -**	**QA =**	**QA +**
IR -	0	0	1	IR -	8	1	4
IR =	9	6	3	IR =	4	2	3
IR +	0	0	0	IR +	4	0	2

We did a full analysis of all the topics, and grouped them according to the factors that hurt/help system performance (Table 9). Our IR4QA paper[2] gives a detailed example for the translation-related issues.

Table 9 Break down of factors that affect performance (CS-CS)

Type	Factors	Topics
bad	Sent-clause extraction model lack the feature "first clause of a sentence"	T43, T52, T55, T64, T65, T68, T69, T85, T323, T324, T340
good	Since clause is shorter, human evaluator is less likely to miss the nugget	T42, T49
good	Sent-clause extraction filter out irrelevant sentence, thus bring relevant sentences up	T41, T46, T47

6. Conclusion

Through our analysis of our CCLQA runs, we identified some factors that can contribute to successful CCLQA systems, such as deep sentence analysis (e.g. syntax parsing), a rich set of features, sequential and tweaked models for answer extraction, clause extraction units and text summarization for answer selection.

Acknowledgment

This work was supported in part by IARPA's Advanced Question Answering for Intelligence (AQUAINT) Program. We thank Eric Riebling for his assistance in corpus preprocessing. We would also like to thank the corpus provider for the Japanese and Chinese corpora.

References

[1] Nyberg, E., R. Frederking, T. Mitamura, M. Bilotti, K. Hannan, L. Hiyakumoto, J. Ko, F. Lin, L. Lita, V. Pedro, and A. Schlaikjer. 2005. JAVELIN I and II systems at TREC 2005. *In Proceedings of TREC'05.*

[2] Lao, N., H. Shima, T. Mitamura and E. Nyberg. 2008. Query Expansion and Machine Translation for Robust Cross-Lingual Information Retrieval. *In Proceedings of NTCIR-7 Workshop*, Japan.

[3] Lin, J., and D. Demner-Fushman. 2006. Methods for Automatically Evaluating Answers to Complex Questions. *Information Retrieval*, 9(5):565-587.

[4] Kor, K.-W. and T.-S. Chua. 2007. Interesting Nuggets and Their Impact on Definitional Question Answering. *In Proceedings of SIGIR 2007*.

[5] Cui, H., M.-Y. Kan, and T.-S. Chua. 2005. Generic Soft Pattern Models for Definitional Question Answering. *In Proceedings of SIGIR 2005*, pages 384–391, New York, NY, USA

[6] Xu, J., A. Licuanan, and R. Weischedel. 2003. TREC 2003 QA at BBN: Answering Definitional Questions. *In TREC '03: Proceedings of the 12th Text REtrieval Conference*, Gaithersburg, Maryland

[7] Harabagiu, S., F. Lacatusu and A. Hickl. 2006. Answering Complex Questions with Random Walk Models. *In Proceedings of SIGIR 2006*.

[8] Mitamura, T., E. Nyberg, H. Shima, T. Kato, T. Mori, C.-Y. Lin, R. Song, C.-J. Lin, T. Sakai, D. Ji and N. Kando. 2008. Overview of the NTCIR-7 ACLIA: Advanced Cross-Lingual Information Access. *In Proceedings of NTCIR-7 Workshop*, Japan.

[9] Shima, H., and T. Mitamura. 2007. JAVELIN III: Answering Non-Factoid Questions in Japanese. *In Proceedings of NTICIR-6 Workshop*, Japan.

[10] Verberne, S. 2007. Paragraph Retrieval for Why-question Answering. *In Proceedings of SIGIR 2007*, Amsterdam, p. 922 (Doctoral Consortium).

[11] Cohen, W.W. 2004. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net

[12] Lafferty, J., A. McCallum and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the ICML2001*.

[13] Sarawagi, S. CRF Package, http://crf.sourceforge.net

[14] Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer. 2006. Online Passive-Aggressive Algorithms. *The Journal of Machine Learning Research*, Volume 7, pages: 551 - 585.

[15] Joachims, T. A. 2001. Statistical Learning Model of Text Classification with Support Vector Machines. *In Proceedings of SIGIR*, ACM.

[16] Chang, C.-C., and C.-J. Lin, LIBSVM: a Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm

[17] Freund, Y. and R. E. Schapire. 1998. Large Margin Classification Using the Perceptron Algorithm. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pages 209-217.

[18] Klein, D., and C. D. Manning. 2002. Conditional Structure versus Conditional Estimation in NLP Models. *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[19] Mccallum, A., D. Freitag, and F. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. *In Proceedings of ICML 2000*.

[20] Carbonell, J. and J. Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In Proceedings of SIGIR 1998.

[21] Goldstein, J., V. Mittal, and J. Carbonell. 2000. Creating and Evaluating Multi-Document Sentence Extract Summaries. *In Proceedings of CIKM* 2000.

[22] Leusch, G., N. Ueffing and H. Ney. 2003. A Novel String-to-string Distance Measure with Applications to Machine Translation Evaluation. *In Proceedings of MT Summit IX*.

[23] S. Blair-Goldensohn, K. R. McKeown and A. H. Schlaikjer. 2003. A Hybrid Approach for Answering Definitional Questions. *Technical Report CUCS-006-03,* Columbia University.

[24] Balage Filho, P.P., V.R. Uzêda, T.A.S. Pardo and M.G.V. Nunes. 2006. Using a Text Summarization System for Monolingual Question Answering. *In Proceedings of CLEF 2006 Workshop,* Alicante, Spain. September 20-22.

[25] Madnani, N., J. Lin, and B. Dorr. 2007. TREC 2007 ciQA Task: University of Maryland. *In Proceedings of TREC*.

[26] Mori, T., M. Nozawa and Y. Asada. 2005. Multi-answer-focused multi-document summarization using a question-answering engine. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 3, pp. 305-320.

[27] Ko, J., L. Si, and E. Nyberg. 2007. A Probabilistic Framework for Answer Selection in Question Answering. *In Proceedings of NAACL-HLT*.

[28] Schlaefer, N., J. Ko, J. Betteridge, G. Sautter, M. Pathak and E. Nyberg. 2007. Semantic Extensions of the Ephyra QA System for TREC 2007. *In Proceedings of TREC*.

[29] Lin, J., D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering. *In Proceedings of INTERACT 2003*, pages 25-32, September 2003, Zurich, Switzerland.

[30] Mitamura, Teruko, Frank Lin, Hideki Shima, Mengqiu Wang, Jeongwoo Ko, Justin Betteridge, Matthew Bilotti, Andrew Schlaikjer and Eric Nyberg. 2007. JAVELIN III: Cross-Lingual Question Answering from Japanese and Chinese Documents, Proceedings of NTCIR-6 Workshop, Tokyo, Japan.