Automated Historical Fact-Checking by Passage Retrieval, Word Statistics, and Virtual Question-Answering

Mio Kobayashi[1] **Ai Ishii**[1] **Chikara Hoshino**[1] **Hiroshi Miyashita**[1]

Takuya Matsuzaki[2]

[1]Nihon Unisys Ltd., Japan
{mio.kobayashi, ai.ishii, chikara.hoshino,
hiroshi.miyashita}@unisys.co.jp
[2]Nagoya University, Japan
matuzaki@nuee.nagoya-u.ac.jp

Abstract

This paper presents a hybrid approach to the verification of statements about historical facts. The test data was collected from the world history examinations in a standardized achievement test for high school students. The data includes various kinds of false statements that were carefully written so as to *deceive* the students while they can be disproven on the basis of the teaching materials. Our system predicts the truth or falsehood of a statement based on text search, word cooccurrence statistics, factoid-style question answering, and temporal relation recognition. These features contribute to the judgement complementarily and achieved the state-of-the-art accuracy.

1 Introduction

The proliferation of social media in the Internet drastically changed the status of traditional journalism, which has been an indispensable building block of modern democracy. News are now produced, propagated, and consumed by people in quite a different way than twenty years ago (Pew Research Center, 2016). The downside is that fake news and hoaxes spread through the social network as quickly as those from trustable sources. A mechanism for *fact-checking*, i.e., finding a support or disproof of a claim in a credible information source, is thus needed as a new social infrastructure.

The sheer amount of the information flow as well as the decentralized nature of the social media calls for support to the fact-checking by information technology (Cohen et al., 2011a,b). Although its full automation seems to be beyond current technology (Vlachos and Riedel, 2014; Hassan et al., 2015), even its partial automation would greatly enhance the power of the current fact-checking services.

Context: ... During the period of the Carolingian dynasty of Francia, the Roman Catholic Church preached that the religious cleansing of sins was necessary in order to achieve salvation after death. ...

Instruction: From (1)-(4) below, choose the one correct sentence concerning events during the 8th century when the kingdom referred to in the underlined portion was established.

Choices:

(1) Pepin destroyed the Kingdom of the Lombards.
(2) Charlemagne repelled the Magyars.
(3) The reign of Emperor Taizong of Tang was called the Kaiyuan era.
(4) The reign of Harun al-Rashid began.

Figure 1: Example of a True-or-False question

As a step towards this direction, we take up the automatic verification of a statement about historical facts against credible information sources. The test statements are collected from the world history examinations in a standardized achievement test for high school students in Japan (the National Center Test for University Admissions, NCTUA). Approximately 60% of the NCTUA world history exams are "True-or-False" questions. A question in this format consists of a paragraph of text that provides the context of the question, an instruction sentence, and four choices (Fig. 1)[1]. One has to choose a correct or incorrect statement from the four choices according to the instruction.

The test statements in the True-or-False questions are thoroughly tuned and checked by the examining board so that they are not too easy nor too difficult for human, and their truth or falsehood can be objectively determined on the basis

[1] The questions are posed in Japanese but we use English in the examples for the sake of readability.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 967–975,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

of the teaching materials written according to the official curriculum guidelines. The automatic verification of these statements hence serves as an idealized but still difficult test-bed for the basic fact-checking technologies.

In previous studies, three main approaches for answering True-or-False questions were presented: passage retrieval (Kano, 2014), conversion to factoid-style question answering (Kanayama et al., 2012), and textual entailment recognition (Tian and Miyao, 2014). In these approaches, the fact-checking task is directly converted to another, existing problem setting. We however show that the test statements in the True-or-False questions have compound characteristics through an analysis of past exams (§3). Thus, direct conversion alone does not suffice for solving this task satisfactorily, because each method is built on its own problem setting, which does not fully cover the variety of the test statements, especially the false ones. In this work, to overcome this difficulty, we attempt to decompose the problems according to our observations of past exams and design a solver that integrates the ideas behind the existing methods as the features of a statistical classifier (§4). Experimental results show that our decomposition of the task of historical fact-checking is successful in that the features work complementarily and the solver achieved the state-of-the-art accuracy (§6). An analysis of the remaining errors indicates a room for further improvement by the incorporation of linguistic and domain-specific knowledge into the system (§7).

Our essential contributions to this problem are as follows.

- Careful observations of past exams were conducted, based on which hypothetical characterizations of the task were formulated. Evidences were then collected to support these hypotheses.
- According to the observed evidences, five features that range over text search, statistics, and logical entailment were designed. They were combined as the features of a classifier and yielded state-of-the-art results on the task.

2 Related Work

Fact-checking can be framed as a question-answering (QA) task in a broad sense. However, it has not been studied as intensively as other QA tasks such as factoid-style question-answering (Ravichandran and Hovy, 2002; Bian et al., 2008; Ferrucci, 2012). Kanayama et al. (2012) proposed to convert a fact-checking question into a set of factoid-style questions. In the conversion, the named entities in a test statement are in turn replaced with an empty slot. The answer, i.e., the most appropriate word that fills the slot, was obtained by an open-domain factoid QA system. They define a confidence score that decreases when the QA system's answer differs from the hidden named entity (i.e., the one replaced with the empty slot). They experimented the idea by *manually* converting the test statements to factoid questions. We follow their idea in designing one of the features. We however fully automatized the conversion and defined another confidence score based on a simple document retrieval system instead of a full-fledged factoid QA system (§4.2.2).

Textual entailment recognition (RTE) (Dagan et al., 2013) has been extensively studied in the field of language processing. RTE can be seen as a quite restricted form of fact-checking where two sentences t and h are given and a system judges whether t is an evidence of (i.e., it entails) h or not. Tian and Miyao (2014) showed the effectiveness of their logic-based RTE system on the True-or-False questions of NCTUA history exams cast in the form of RTE (i.e., a test statement *and* an evidence sentence are given to the system).

Recent effort pursued a more realistic task setting for RTE, in which a system is given a sentence h and a large number of candidates of its evidence $\{t_i\}$ that are drawn from a document collection in advance. The system tests the entailment relation between each of t_is and h (Bentivogli et al., 2010, 2011). In the RITE-VAL shared task in NTCIR-2014 conference (Matsuyoshi et al., 2014), the participating RTE systems were evaluated both in the traditional RTE task setting and one that fully integrates the document retrieval and entailment recognition (i.e., only h and a document collection are given to the systems). The test sentences (i.e., hs) included those taken from NCTUA world history exams and hence the latter task setting is close to ours. The performance degradation between the two task settings was around 14% (absolute) for the case of the best-performing system. It indicates the difficulty of our task setting of the historical fact-checking.

In a series of recent papers, elementary science questions are used as a benchmark of AI systems (Khot et al., 2015; Jansen et al., 2016; Clark et al., 2016; Khashabi et al., 2016). The questions, all in the form of multiple-choice questions, were collected from 4th grade science tests. In the majority of the questions, the choices are nouns rather than sentences as in the following example taken from (Khashabi et al., 2016):

> Q. In New York State, the longest period of daylight occurs during which month? (A) December (B) June (C) March (D) September

The majority of them can hence be regarded as a factoid-style question with hints (i.e., the answer is one of the four). Clark et al. (2016) and Khashabi et al. (2016) demonstrated that the system performance was boosted by multi-step inference that combines, e.g., taxonomic knowledge ("N.Y. is in the northern hemisphere") and general law ("The summer solstice in the northern hemisphere is in June"). A special kind of logical relation, temporal inclusion, is considered in our system (§4.3). The intention is however on the detection of the falsity of a statement that is not in a temporal inclusion relation with an evidence sentence. The feature based on the conversion to factoid questions is also designed for the detection of falsehood by finding a counter-evidence. The different orientations, i.e., proof of a scientific fact vs. disproof of a historical non-fact, reflect the different natures of the problems.

3 Observation of Task

We examined past True-or-False questions prior to implementing the solver. The observation targets were the NCTUA world history exams 2005, 2007, 2009, 2011, and 2013s (supplementary exam). We used four sets of high school textbooks of world history and Wikipedia as knowledge resources.

From the observations, we formulated three hypotheses as follows. First, to verify most of the test statements (i.e., the choices), it is not necessary to gather several evidence sentences across different paragraphs in the resources; usually there is sufficient information in a local portion, such as a paragraph or a sentence in a knowledge resource. This is a natural consequence of the fact that most of the test statements describe a single historical event. Second, the knowledge resources include a more detailed time expression than the questions, e.g., "1453" as compared to "15th Century," and "1945" as compared to "1940s." Third, the falsity of many incorrect statements is attributed to a single named entity (NE) or time expression in them. For example, Choice (2) in Fig. 1, "Charlemagne repelled the Magyars." is an incorrect statement created by changing "Avars" to "Magyars" in a correct sentence "Charlemagne repelled the Avars."

Knowledge resource	Count (ratio)
Textbook	111/137 (0.81)
Wikipedia	118/137 (0.86)
Textbook + Wikipedia	129/137 (0.94)

Table 1: Ratio of correct statements that can be evidenced by a single paragraph in the knowledge resource

	Count (ratio)
NE change	163 / 275 (0.59)
Time change	47 / 275 (0.17)
NE or Time change	210 / 275 (0.76)

Table 2: Ratio of incorrect statements in which one named entity or time expression is the reason of the falsity

To support these hypotheses, we gathered evidences from past exams. First, we examined the correct statements (137 in total) to determine whether: (1) a single paragraph in the knowledge resources includes all the NEs in the statement and (2) the knowledge resources include a more detailed time expression than the statement. Table 1 shows that, in most cases, a single paragraph includes sufficient information to allow the solver to verify the truth of a statement. Among the 137 correct statements, 48 of them included a time expression. The knowledge resources provided a more detailed time for 45 out of these 48 statements (94%). It is thus important to resolve the level of detail of the time expressions. Next, we counted the incorrect statements (275 in total) which can be turned into a correct one by changing one NE or time expression in it. Table 2 shows that to detect the falsity of an incorrect statement, detection of the changed NE is crucial.

4 Features and their Combination

Based on the above observations, we designed the following five features to score the confidence of truth. This section describes them in turn and explains how they are combined as the features of a statistical classifier.

These features are defined using several statistics collected on a set of documents. We created three document sets, D_s, D_p, and D_m, all from Wikipedia and four high school textbooks of world history, as follows. We first segmented the Wikipedia pages and the textbooks in two ways: one into a set of sentences D_s and the other into a set of paragraphs D_p. D_m is the union of D_s and D_p.

4.1 Text Search Feature

The observations revealed that, in most cases, the NEs in a correct statement are fully included in one paragraph or one sentence in the knowledge resources. The text search feature of the statement S is defined as the number of documents in D_m which include all NEs and content nouns in S. The solver expects that the greater the number is, the more likely the statement is true.

4.2 Statistical Features

The observations showed that an incorrect statement is created mainly by changing one NE in a correct sentence. To detect such a conversion, the solver estimates the strength of relatedness among the NEs in the statement. The solver uses two statistical features, which are respectively defined using global and local statistics collected on the document set.

4.2.1 Pointwise Mutual Information (PMI) Feature

The first statistical feature is PMI (Church and Hanks, 1989), which is defined for a pair of words as,

$$pmi(w_1, w_2) \equiv \log \frac{p(w_1, w_2|D_s)}{p(w_1|D_s)p(w_2|D_s)},$$

where $p(w_i|D_s)$ denotes the probability of observing the word w_i in a sentence that is randomly chosen from D_s and $p(w_1, w_2|D_s)$ denotes the probability of observing both w_1 and w_2 in a randomly chosen sentence. A low PMI score indicates the independence of the two words. The solver expects that a low PMI score indicates that two NEs are not related to each other and suggests that the statement is incorrect.

The solver calculates PMI for the pairs of an NE and the subsequent NE or a content word that appears before the subsequent NE in the statement, because the two words positioned close to each other tend to have strong relation. We write $\mathrm{WP}(S)$ for the set of such pairs of words in S excluding the pairs of synonymous words. The final PMI feature of the statement S is defined by the average of $pmi(w_i, w_j)$.

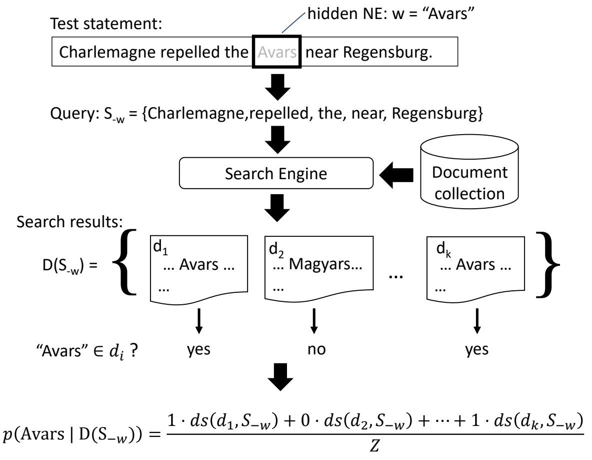
Test statement:
hidden NE: w = "Avars"
Charlemagne repelled the Avars near Regensburg.
Query: S-w = {Charlemagne,repelled, the, near, Regensburg}
Search Engine
Document collection
Search results:
D(S-w) =
d1 ... Avars ... ...
d2 ... Magyars... ...
...
dk ... Avars ... ...
"Avars" ∈ di ?
yes
no
yes
p(Avars | D(S−w)) = (1 · ds(d1,S−w) + 0 · ds(d2,S−w) + ⋯ + 1 · ds(dk,S−w)) / Z

Figure 2: Calculation of the probability $p(w|D(S_{-w}))$ of finding w in the search results $D(S_{-w})$

$$\overline{pmi}(S) = \frac{1}{|\mathrm{WP}(S)|} \sum_{(w_i,w_j)\in\mathrm{WP}(S)} pmi(w_i, w_j).$$

4.2.2 Virtual Question Answering (VQA) Feature

The second feature is based on the result of a search that approximates the process of factoid QA. It directly attempts to detect the changed word in an incorrect statement (Fig. 2). The solver hides each NE w in a statement S in order and makes a VQA query S_{-w}, which is the set of words in S excluding w. If the statement is correct, the hidden word is expected to be found in the search results of the query S_{-w} at high probability.

For a hidden NE w, we first calculate the ratio $vqa(w)$ of the probabilities of finding w in the search results $D(S_{-w})$ of the query S_{-w} and in a randomly chosen document in the collection D_m:

$$vqa(w) = \log \frac{p(w|D(S_{-w}))}{p(w|D_m)}.$$

In the numerator, we consider the probability of finding w in a document $d \in D(S_{-w})$ that is sampled according to the confidence on the search result d against the query S_{-w}, rather than assuming a uniform distribution on $D(S_{-w})$:

$$p(w|D(S_{-w})) = \sum_{d \in D(S_{-w})} [w \in d] \cdot p(d|D(S_{-w})), \qquad (1)$$

where $[w \in d]$ is the binary indicator function that takes value 1 if d includes w, and 0 otherwise. The confidence factor $p(d|D(S_{-w}))$ is assumed to be proportional to a document score $ds(d, S_{-w})$ and we only consider top-k search results. That is, letting d_i denote the document ranked i-th in the search results according to $ds(d, S_{-w})$, we assume

$$p(d_i|D(S_{-w})) = \begin{cases} ds(d_i, S_{-w}) \cdot Z^{-1} & (1 \le i \le k) \\ 0 & (k < i) \end{cases} \qquad (2)$$

where $Z = \sum_{j=1}^{k} ds(d_j, S_{-w})$ is the normalization factor. From (1) and (2), we have

$$p(w|D(S_{-w})) = \sum_{\substack{i=1 \\ w \in d_i}}^{k} \frac{ds(d_i, S_{-w})}{\sum_{j=1}^{k} ds(d_j, S_{-w})}.$$

We set $k = 30$ in our experiments.

The document score ds is defined by Term Frequency-Inverse Document Frequency (TF-IDF), which is written as

$$ds(d, S_{-w}) = \frac{1}{\ell(d)} \sum_{w' \in S_{-w}} tf(w', d) \cdot idf(w'),$$

where $tf(w', d)$ is the frequency of the word w' in the document d, $idf(w')$ is the inverse document frequency of the word w', and $\ell(d)$ is the length of d.

The VQA feature uses document-local statistics (except for IDF) and counts only in the top-k search results. However, in this feature, all the query words jointly contribute through the document score ds, in contrast to the case of PMI where only the pairwise relations are considered. The final VQA feature is defined as the average of $vqa(w)$.

$$\overline{vqa}(S) = \frac{1}{|\mathrm{NE}(S)|} \sum_{w \in \mathrm{NE}(S)} vqa(w),$$

where $\mathrm{NE}(S)$ is the set of NEs in the statement S.

4.2.3 Length Feature

We also use the length of the statement (number of words) as a feature. PMI and VQA features of a long statement tend to have low values regardless of the correctness of the statement. The length feature adjusts this bias.

4.3 Time Feature (Logical Entailment)

The observations revealed that the level of detail of the time expressions in the choice sentences differs from that in the knowledge resources. The time information is a key factor of historical events. Therefore, the solver needs a more rigorous inference about temporal relations than about the matching of other NEs. We implemented a module that logically determines the inclusion relation between two time expressions. The time expressions in the statements and the knowledge resources are extracted and converted to ranges of date (e.g., “19th century” → 1801-01-01 ... 1900-12-31) by NormalizedNumexp[2]. If the range of a time expression in a statement includes one in the knowledge resources, they are judged as “matched”. The solver hides the time expression t from the statement S and makes the VQA query S_{-t}. The time feature of the statement S is defined as the number of documents in the top-k search results of query S_{-t} ($k = 30$) that include a time expression that matches the hidden time expression t.

4.4 Combination of Features (Machine Learning)

The solver combines the above five features using statistical binary classifiers. In our settings, N training samples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ are given, where x_i is a five-dimensional feature vector and $y_i \in \{1, 0\}$ indicates the truth or falsehood of the sample. We used the “scikit-learn” toolkit[3] and created an ensemble of three classifiers, by simply averaging their [0, 1] probabilistic outputs to reduce variance of each classifier (Pedro, 2012). As the classifiers, we used logistic regression, gradient boosting classifier, and support vector machine. The hyperparameters of each classifier are determined by cross validation.

5 Resources and Common Modules

This section describes additional modules that were used in the experiments described in §6.

[2] https://github.com/nullnull/normalizeNumexp

[3] http://scikit-learn.org/stable/

Dataset	# test statements	%correct	%incorrect
DEV	412	33.3%	66.7%
TEST	1112	35.3%	64.7%

Table 3: Size of development and test data

5.1 Custom Dictionary

We use a named entity dictionary and a synonym dictionary, both of which were manually compiled based on textbooks and Wikipedia. The named entity dictionary was created by mainly using the index of textbooks. In the dictionary, approximately 20,000 NEs are categorized into various classes (time, person, etc.) by human experts. The synonym dictionary was created based on Wikipedia redirect and bracketed expressions after NEs (e.g., "Charlemagne (Charles I)"). Additionally, the solver uses Nihongo Goi-Taikei[4], a Japanese thesaurus, to discriminate NEs from common nouns.

5.2 Retrieval Module

The retrieval module of the solver is based on Apache Solr[5]. We used the Solr defaults (TFIDF-weighted cosine similarity) and the Kuromoji Japanese morphological analyzer[6] to tokenize Japanese sentences. As mentioned in §4, all knowledge resources are indexed at overlapping levels of the sentence and the paragraph, and retrieval is executed across fields of both levels by means of the ExtendedDisMax Query Parser[7].

5.3 Matching of Words

When two words are compared in the solver, some suffixes are ignored to absorb orthographical variants (e.g., "Japan" and "Japanese" are considered to be the same). The suffix list is made from high frequency morphemes (Okita and Liu, 2014). We examined the frequency of morphemes in the textbooks, and then from the top, if the morpheme is a suffix of NE, we added to the list.

Additionally, if a word w in a question has a synonym in a document retrieved from the knowledge resources, the word w is considered to be included in the document.

5.4 Complementing the Lack of Information

The truth or falsehood of a choice sentence is often indeterminable without the information provided in the context and the instruction. For instance, (1), (3), and (4) in Fig. 3 all describe historical facts but the condition in the instruction, "in the 20th century," turns (1) and (4) to false since they happened in the 19th century. Meanwhile, the context and the instruction also include irrelevant information that does not affect the truth of the choices. For instance, the underlined portion, "Iberian Peninsula," is redundant since the instruction asks more restrictively about "Spain." Furthermore, "the Middle Ages" in the context is not relevant to any of the choices.

Context: Coexistence between Christians and Muslims was seen on the Iberian Peninsula in the Middle Ages, ...

Instruction: From (1)-(4) below, choose the most appropriate sentence that describes the history of Spain in the 20th century related to the underlined portion.

Choices:

(1) The French army suffered in guerrilla warfare in Spain.
(2) In the Spanish Civil War, Germany and Italy maintained a policy of non-intervention.
(3) Franco established a dictatorial regime.
(4) The Philippines were seized from it by the U.S.A.

Figure 3: Question 31 in the 2011 data set

The instruction tends to include a condition that applies to all choices, such as the location and the time of the historical events described in them. The context usually provides relevant condition only when it is explicitly indicated so.

We utilize these observations as well as the category of the NEs to extract only relevant keywords from the instruction and the context. First, the location names and time expressions are extracted from the instruction if a choice includes no such phrases. The NEs in the underlined portion of the context are then extracted if the instruction includes none of the phrases in a pre-defined set of cue phrases, such as "related to," that indicate the context is not so relevant to the determination of the truth of the choices. Finally, among the NEs extracted from the context, we discard those categorized as an abstract concept in the NE dictionary, such as "social phenomena" and "social role." For the choice (1) in Fig. 3, "the 20th century" in the instruction is extracted since (1) does not include a time expression but "Iberian Peninsula" in the context is not extracted since the instruction refers to the underlined portion using the cue phrase "related to."

[4] http://www.iwanami.co.jp/hotnews/GoiTaikei/
[5] http://lucene.apache.org/solr/
[6] http://www.atilika.org/
[7] https://cwiki.apache.org/confluence/display/solr/The+Extended+DisMax+Query+Parser

Features	DEV		TEST	
	Binary T/F	4-way	Binary T/F	4-way
All	80.8%	75.7%	74.2%	68.0%
-Text search	77.7%	66.0%	70.5%	60.1%
-PMI	79.6%	74.8%	73.7%	66.2%
-VQA	74.0%	64.1%	71.2%	60.8%
-Time	79.6%	75.7%	73.4%	66.9%
-Length	80.6%	75.7%	73.8%	67.3%

Table 4: Feature ablation study

Features	DEV		TEST	
	Binary T/F	4-way	Binary T/F	4-way
Text search	73.8%	58.3%	71.0%	52.5%
PMI	67.7%	53.4%	66.3%	45.7%
VQA	74.8%	58.3%	70.8%	53.2%
Time	68.2%	35.0%	65.3%	28.4%
Length	66.5%	33.0%	65.3%	23.7%

Table 5: Accuracy with only one feature

6 Experiments

6.1 Experimental Setup

We exhaustively extracted the True-or-False questions from past NCTUA world history exams held from 2005 to 2015 and evaluated the accuracy of the true-or-false (T/F) binary predictions individually made on each of the choice sentences[8]. The data was divided into two disjoint subsets, DEV and TEST. DEV consists of the questions used in the preliminary analysis described in §3. TEST consists of the rest of the questions. Table 3 provides the number of the test statements (i.e., the choice sentences) and the distributions of the correct and incorrect statements. Approximately 20% of the questions ask to choose a false statement in four choices, which include three correct statements. For reference, we also report the accuracy on the 4-way multiple-choice questions. The answer to a multiple-choice question is the choice on which the ensemble of the classifiers yielded the maximum or minimum score.

For the evaluation, we adopted cross validation. DEV and TEST were divided into 20 subsets, each of which was taken from the questions in the same exam. We applied 20-fold cross validation on the subsets. We summarized the results with DEV and TEST respectively.

6.2 Experimental Results

To evaluate the importance of each feature, we tested two feature combination patterns. In the first pattern, the classifiers were trained excluding one feature (Table 4) and in the second one the classifiers were trained with only one feature (Table 5). These results show that the VQA and Text search features are more important than the others. The highest T/F judgement accuracy, 80.8% on DEV and 74.2% on TEST, was obtained using all the features. On the DEV set, the ablation of one of the five features resulted in a loss of 0.2-6.8 points in the T/F judgement accuracy and 0.0-11.6 points in the 4-way multiple-choice accuracy (Table 4). It suggests that the combination of the features is more effective for comparing the confidence on the truth of four statements rather than for the T/F judgement on a single statement. Comparison of the results in Table 5 with 'All' in Table 4 further supports it; the effect of combining the five features compared with the result by a single feature is far more evident in the accuracy on the multiple-choice format. These results show the five features worked complementarily and validates our decomposition of the task into the five features.

System	T/F binary acc.	4-way acc.
Kanayama et al. (2012)	79% (73/92)	65% (15/23)
This paper (VQA only)	73% (67/92)	52% (12/23)
This paper (all features)	84% (77/92)	83% (19/23)

Table 6: Comparison with manual question conversion on NCTUA 2007 questions

Table 6 presents a comparison with Kanayama et al. (2012)'s result based on the conversion of T/F judgement to factoid-style questions. The test questions were taken from NCTUA 2007 exam. This comparison is not strict in a few regards. First, Kanayama et al. used English translations of the questions while we used the original questions in Japanese. Second, they manually supplied the choice sentences with necessary information extracted from the instruction and the context. Finally, they converted the choice sentences to factoid-style questions also manually, while our system is fully automated. The results of VQA feature are slightly worse than Kanayama et al.'s which is based on a similar idea. The addition of the other four features boosted the accuracy and it surpassed their result.

Finally, Table 7 presents the accuracy of our system in comparison with previously reported results by fully automatic systems (Shibuki et al., 2014, 2016). We compared the accuracy on the

[8] Although the instruction sentences indicate either (i) only one of the choices is correct ("choose the correct one") or (ii) only one of them is incorrect ("choose the incorrect one"), our solver does not utilize this information in any form when it makes binary T/F prediction.

System	4-way acc. NCTUA 2007
This paper	83% (19/23)
Okita and Liu (2014)	74% (17/23)
Kano (2014)	57% (13/23)
Sakamoto et al. (2014)	52% (12/23)

System	4-way acc. NCTUA 2011
This paper	90% (18/20)
Kobayashi et al. (2016)	80% (16/20)
Takada et al. (2016)	60% (12/20)
Sakamoto et al. (2016)	60% (12/20)

Table 7: Comparison with previous automatic systems

True-or-False questions in the 4-way multiple-choice format. Our system achieved a higher accuracy than the best previous results on NTCUA 2007 and 2011 exams.

7 Error Analysis

We now show some examples that cannot be solved by our current approach and describe the cause of the errors.

Antonym of Verbs (10 sentences) Some incorrect statements in the choices are created by changing a verb to its antonym. For example, in Question 8 in the 2009 exam, the falsity of the sentence "The Agricultural Adjustment Act (AAA) resulted in the prices of agricultural produce being lowered," is attributed to the verb "lowered" because the sentence becomes correct if we replace "lowered" with "raised." To properly recognize such false sentences, we need to utilize lexical knowledge about the antonymy and synonymy relations among verbs.

Semantic Roles (five sentences) Many historical events involve two or more participants. For example, in Question 3 in the 2007 exam, the sentence "The Almohad Caliphate, which advanced into the Iberian Peninsula, was overthrown by the Almoravid dynasty." includes the two participants, "The Almohad Caliphate" and "the Almoravid dynasty." The sentence is incorrect because the truth was "The Almoravid dynasty was overthrown by the Almohad Caliphate." To detect this kind of falsity, we need to recognize the semantic roles (e.g., agent and patient) of the participants in the event denoted by the verb. It is beyond the expressiveness of the VQA and PMI features that are largely based on word cooccurrence.

Indirect Description of Time (four sentences) In Question 15 in the 2007 exam, the instruction includes the following phrase: "choose the one term that correctly describes the religion that was established after the time of Genghis Khan." The current system cannot extract any temporal information from "after the time of Genghis Khan," which is equivalent to "after the *death* of Genghis Khan" and thus to "after 1227." To this end, we need to analyze the combination of the temporal expression (e.g., the time before/after/during of X) and the entity type (e.g., X: Person) and utilize domain-knowledge such as birth/death or establishment/abolishment years of historical entities.

8 Conclusion and Future Work

As a step towards the goal of automated fact-checking, we have worked on the task of true-or-false judgement on the statements about historical facts. We scrutinized the characteristics of the task and designed five confidence metrics according to the observations, which are integrated as the features in a statistical classifier. Experimental results showed that the five features complementarily contributed to the discrimination between a true statement and a false one. Our system achieved the state-of-the-art accuracy on a few datasets. An analysis of the remaining errors indicated a room for improvement by the incorporation of linguistic knowledge such as antonymy of verbs and semantic roles of the events, and extraction of temporal information based on linguistic patterns and domain-knowledge.

Acknowledgments

This research was supported by Todai Robot Project at National Institute of Informatics. We are gratefully acknowledge Tokyo Shoseki Co., Ltd. and Yamakawa Shuppansha Ltd. for providing the textbook data.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In *TAC*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *TAC*.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *WWW*. pages 467–476.

K. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *ACL-89*. pages 76–83.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI-2016*. pages 2580–2586.

Sarah Cohen, James T. Hamilton, and Fred Turner. 2011a. Computational journalism. *Commun. ACM* 54(10):66–71.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011b. Computational journalism: A call to arms to database researchers. In *CIDR*. pages 148–151.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers.

David A. Ferrucci. 2012. Introduction to “this is watson”. *IBM Journal of Research and Development* 56(3.4):1:1–1:15.

Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proc. of the 2015 Computation+Journalism Symposium*.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *COLING*. pages 2956–2965.

Hiroshi Kanayama, Yusuke Miyao, and John Prager. 2012. Answering yes/no questions via question inversion. In *COLING-2012*. pages 1377–1392.

Yoshinobu Kano. 2014. Solving history exam by keyword distribution: KJP. In *NTCIR-11*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. In *IJCAI*. pages 1145–1152.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. 2015. Exploring markov logic networks for question answering. In *EMNLP*. pages 685–694.

Mio Kobayashi, Hiroshi Miyashita, Ai Ishii, and Chikara Hoshino. 2016. NUL system at QA Lab-2 task. In *NTCIR-12*.

Suguru Matsuyoshi, Yusuke Miyao, Tomohide Shibata, Chuan-Jie Lin, Cheng-Wei Shih, Yotaro Watanabe, and Teruko Mitamura. 2014. Overview of the NTCIR-11 recognizing inference in text and validation (RITE-VAL) task. In *NTCIR-11*.

Tsuyoshi Okita and Qun Liu. 2014. The question answering system of DCUMT in NTCIR-11 QA Lab. In *NTCIR-11*.

Domingos Pedro. 2012. A few useful things to know about machine learning. In *Commun. of the ACM*. volume 55(10), pages 78–87.

Pew Research Center. 2016. News use across social media platforms 2016.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *ACL-2002*. pages 41–47.

Kotaro Sakamoto, Madoka Ishioroshi, Hyogo Matsui, Takahisa Jin, Fuyuki Wada, Shu Nakayama, Hideyuki Shibuki, Tatsunori Mori, and Noriko Kando. 2016. Forst: Question answering system for second-stage examinations at NTCIR-12 QA Lab-2 task. In *NTCIR-12*.

Kotaro Sakamoto, Hyogo Matsui, Eisuke Matsunaga, Takahisa Jin, Hideyuki Shibuki, Tatsunori Mori, Madoka Ishioroshi, and Noriko Kando. 2014. Forst: Question answering system using basic element at NTCIR-11 QA-Lab task. In *NTCIR-11*.

Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. 2016. Task overview for NTCIR-12 QA Lab-2. In *NTCIR-12*.

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab task. In *NTCIR-11*.

Takuma Takada, Takuya Imagawa, Takuya Matsuzaki, and Satoshi Sato. 2016. SML question-answering system for world history essay and multiple-choice exams at NTCIR-12 QA Lab-2. In *NTCIR-12*.

Ran Tian and Yusuke Miyao. 2014. Answering center-exam questions on history by textual inference. In *Proceedings of the 28th Annual Conference of the Japanese Society for Artificial Intelligence*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. ACL 2014 Workshop on Language Technologies and Computational Social Science*. pages 18–22.