Automatic Generation of Japanese Question-Answering Pairs

Hiroki Tanioka*1 Kaoru Kimura*1
Kazuma Takaoka*2 Ryohei Nakatani*2 Yoshitaka Uchida*2

*1 Tokushima University, Tokushima, Japan
{tanioka.hiroki, c501506035}@tokushima-u.ac.jp

*2 Works Applications, Tokushima, Japan
{takaoka_k, nakatani_r, uchida_yo}@worksap.co.jp

Faculty of Engineering
Tokushima University

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

Introduction

What's the goal?

Question-answering pairs are needed for a lot of question-answering and chatbot systems. Because those systems are based on search engine.
If many and high-quality question-answering pairs are indexed on the search engine, the question-answering system is expected to make a response to users.

Japanese Corpus Needed

The question-answering system needs lots of sophisticated question-answering corpora. There are already large- scale English question answer corpora available.
However, there are no large-scale Japanese question-answer corpus. Therefore, we need more Japanese question-answering corpora.

Machine Learning!!

A low-cost approach is needed for generating question-answering corpus. Furthermore, various types of question-answer corpora are needed for each domain.
To make the index quickly, a machine learning technique can be used for generating question-answering pairs automatically.

It's a challenge to generate question sentences automatically.

Approach

Preparing Question-Answering Pairs

Step 1. Choose a category in the following 10 categories.
Step 2. Choose 10 topics in the category.
Step 3. Make 6 types of questions in each topic.
Step 4. Describe 1 or more answers for each question.

Category	Number of Themes	Number of Questions	Number of Answers
学術(Academia)	10	100	100
技術(Technology)	10	124	156
自然(Nature)	11	127	146
社会(Society)	11	108	132
地理(Geography)	10	102	100
人間(Human)	10	73	74
文化(Culture)	10	136	138
歴史(History)	10	54	61
マニュアル(Manual)	10	74	74
シラバス(Syllabus)	15	120	120

Total
Questions: 1,018
Answers: 1,101

Extracting Candidate Sentences

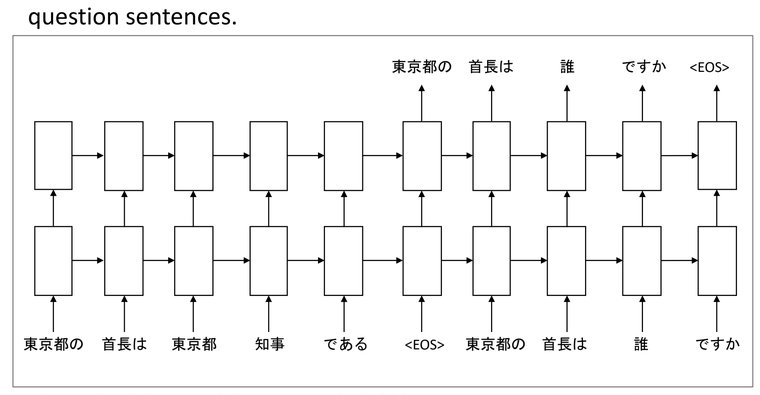
Extracting candidate sentences using Support Vector Machine (SVM) with Sudachi. The accuracy is 90.1%, and the precision is 41.4%.

Type	SVM	Human	Target Sentence
True Positive	✓	✓	東京都の首長は、東京都知事である。 The head of Tokyo is the governor of Tokyo.
	✓	✓	地震に対して、地殻が非常にゆっくりとずれ動く現象を地殻変動と呼ぶ。 Crustal deformation is a phenomenon in which the crust moves very slowly relative to an earthquake.
False Negative		✓	魚料理や肉料理などの主菜など、イタリア料理のコースの流れがサブタイトルになっている。 A flow of course of Italian cuisine such as main dish such as fish dish and meat dish is subtitle.
		✓	1 メガパーセクは 326 万光年。 1 Mega persec is 3.26 million light-years.
False Positive	✓		2012 年現在、国際的な本初子午線として IERS 基準子午線が使用されている。 As of 2012, the IERS standard meridian is used as the international prime meridian.
	✓		東京都の議決機関は東京都議会である。 The decision-making body in Tokyo is the Tokyo Metropolitan Assembly.
True Negative			ツンドラ・タウン（Tundratown） 寒冷地域の動物たちが暮らすエリア。 Tundratown, an area is where cold region animals live.
			ディズニーのアニメーション映画で 10 億ドルを突破したのは、『トイ・ストーリー3』、『アナと雪の女王』に次いで史上 3 番目である。 It was the third one in history, after "Toy Story 3" and "Anna and the Snow Queen" that broke through $ 1 billion in Disney's animated film.

* For Sudachi, analysing mode A, with the full version dictionary. For SVM, the kernel of SVM is RBF. The parameters are "c=1000 -w0 10 -w1 1". Train data contains 5,000 sentences including 188 sentences chosen manually by human. Test data contains 2,000 sentences.

Learning Question Sentences

Attention Sequence to Sequence Model for learning to generate question sentences.



* Referenced code: https://github.com/Gin04gh/nlp/blob/master/AttentionSeq2Seq.ipynb

Evaluation

BLEU and METEOR

BLEU is an algorithm for the evaluation machine translation output, based on matching reference cooccurrence in N-grams. Here, $N = 4$ and $BP_{BLEU} = 1$.

$$p_n = \frac{\sum_i S_i}{\sum_i T_i},$$

$$BLEU = BP_{BLEU} * \exp(\frac{1}{N}\sum_{n=1}^{N} \log p_n).$$

METEOR is a metric for the evaluation machine translation output, based on a F-mean score with precision P and recall R. $\alpha = 0.9$, $\beta = 3$ and $\gamma = 0.5$.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R},$$

$$METEOR = F_{mean} \cdot (1 - \gamma(\frac{c}{u})^{\beta}).$$

* c means the number of words in both a generated sentence and a correct sentence, and u means the number of words in the generated sentence, with matching reference words in the correct sentence.
* S_i is the number of N-grams in a generated sentence i with the matching reference cooccurrence in a correct sentence. T_i is the number of N-grams in a generated sentence i. BP_{BLEU} means a penalty when a sentence is too shorter than a correct sentence.

PER and DER

Position independent word Error Rate (PER) is an evaluation metrics, based on sentence similarity error.

$$PER = 1 - \frac{\sum_i d_i}{\sum_i m_i}$$

Dependency Error Rate (DER) is one of metrics using a dependency parser for evaluating Japanese fluency.

$$DER = 1 - \frac{\sum_i g_i}{\sum_i f_i}$$

Where g means the number of dependency pairs in a generated sentence, with matching reference dependency pairs in a correct sentence, and f means the number of dependency pairs in a correct sentence.

Here, CaboCha is employed for dependency parser. The following example includes chunks and tokens. Chunks are underlined.

Parsed sentence:
「東京都の首長は、東京都知事である。」
（The head of Tokyo is the governor of Tokyo. ）

Dependency pairs:

東京 (Tokyo)	→ 都 (city)
都 (city)	→ の (of)
東京都の (of Tokyo city)	→ 首長は、 (The head,)
首長 (head)	→ は <subject>
は <subject>	→ 、 (,)
首長は、 (The head,)	→ 東京都知事である。 (is the governor of Tokyo.)
東京 (Tokyo)	→ 都知事 (the governor of Tokyo)
都知事 (the governor of Toyo)	→ で (is)
で (is)	→ ある (is)
ある (is)	→ 。 (.)

Candidate sentence	Generated question sentence	Correct Sentence	BLEU	METEOR	PER*	DER*	Score
世界遺産：世界遺産は、「顕著な普遍的価値」を有する文化遺産や自然遺産などであり、1972 年に成立した世界遺産条約に基づき、世界遺産リストに登録された物件を指す。 World Heritage: The World Heritage are cultural heritage and natural heritage which have "remarkable universal value," and it refers to the property registered in the World Heritage list based on the World Heritage Convention established in 1972.	世界遺産は何ですか？ What is the World Heritage Site?	世界遺産は、「顕著な普遍的価値」を有する文化遺産や自然遺産のことですか？ Is the world heritage a cultural heritage or natural heritage having "Remarkable Universal Value"?	0.275	0.353	0.273	0.182	-1
世界遺産：世界遺産は、政府間委員会である世界遺産委員会の審議を経て決定される。 World Heritage Site: The World Heritage Site is decided after deliberation by the World Heritage Committee which is an intergovernmental committee.	世界遺産は政府の委員会ですか？ Is the World Heritage a government committee?	世界遺産は、どこの審議を経て決定されますか？ Which deliberation is the World Heritage decided after?	0.156	0.309	0.375	0.200	0
東京都：東京都の首長は、東京都知事である。 Tokyo: The head of Tokyo is the governor of Tokyo.	東京都の首長は何ですか？ What is the head of Tokyo?	東京都の首長を何といいますか？ What do you call the head of Tokyo?	0.181	0.366	0.600	0.300	2
地震：地震に対して、地殻が非常にゆっくりとずれ動く現象を地殻変動と呼ぶ。 Earthquake: A phenomenon in which the crust moves very slowly against earthquake is called crustal deformation.	地震が地殻には何を出すに何と言いますか？ What does the earthquake say to what to put out on the crust?	地震に対して，地殻が非常にゆっくりとずれ動く現象を何と呼ぶか？ What do you call a phenomenon that the crust moves very slowly against earthquakes?	**0.041**	0.325	0.526	**0.063**	-1
本初子午線：「本初」とは「最初・首位」という意味である。 Prime meridian: "prime" means "the first and the first place".	本初子午線の本初たる最初は何を出す意味ですか？ What does the first as the prime of prime meridian mean to put out?	本初子午線の「本初」とは「最初・首位」をという意味ですか？ Does "Prime" of prime meridian mean "the first and the first place"?	0.288	0.377	0.550	0.222	3
動物園：水中の動物を中心として扱うものは特に水族館とされ、動物園の特殊な形態としてサファリパークや移動動物園、鳥類園、クマ牧場などがある。 Zoo: The ones that keep mainly with animals underwater are considered to be particularly aquariums, and special forms of zoos include safari parks, mobile zoos, bird gardens, bear pastures and on	水中の動物を中心として、形態のサファリパークが移動動物園されているか？ Mainly underwater animals, are safari parks of the form mobile zoos?	一般に陸上の動物を中心として扱う動物園に対して，水中の動物を中心として扱うものは特に何と言われるか？ In general, what are especially those keep mainly underwater animals against zoos keep mainly land animals?	0.319	0.280	0.382	0.214	0
日本：連合国の指令のもと、国制の改革が進められ、大日本帝国憲法の改正手続きによって日本国憲法を制定し、1947 年発効の同憲法によって「国民主権」「平和主義」「基本的人権の尊重」の三大原則を確立した。 Japan: Under the direction of the Allied Powers, reform of the national system was promoted, the Constitution of Japan was enacted by the amendment procedure of the Constitution of the Japan Imperial Empire, and the Constitution which came into effect in 1947, entitled National Sovereignty, Pacifism, and Respect for Human Rights	指令のもとの国制の改革を大日本帝国憲法していますか？ Does the Constitution of the Japan Imperial Empire reform the national system under the directive?	日本国憲法が施行されたのは何年ですか？ What year was the Constitution of Japan enforced?	**0.029**	0.119	0.267	**0.077**	-3

* PER* = 1 – PER, and DER* is 1 – DER.

The average scores of evaluation results to 196 candidate sentences.

Metrics	Score
BLEU	0.095
METEOR	0.235
PER	0.388
DER	0.137

* BLEU score is 2-gram based. METEOR score did not employ WordNet. PER

Discussion

For qualitative evaluation, the generated sentences have grammatical errors and unnatural particles in Japanese.

A blue boxed sentence is scored **good** in each metric without DER.

A red boxed sentence is scored **wrong** in each metric without DER.

Score is calculated by counting good square (+1) and wrong square (-1) for each sentence without DER. Then, **DER is fit to Score.**

Conclusion

- Japanese question sentences are generated from labeled corpus.
- Extracting candidate sentences is high accuracy 90% using SVM. However, Generating question sentences is still low scores in each metric.
- DER is reasonable metric, because DER has larger reduction rate than PER to grammatical error. Also, DER shows a average score.
- Hence, further research to increase generating accuracy and detect answers is necessary with other examples and machine learning models.

What is JQAC Library?

Japanese Question-Answering Corpus (JQAC) is a dataset, consisting of question-answering pairs, which is manually made by university students on a set of Japanese Wikipedia articles and some public documents.
(distributed under the CC BY-SA 4.0 license):

JQAC Library. (2018). Japanese Question-Answering Corpus.
https://taniokah.github.io/jqac/