Impact Objectives

- Develop innovative technologies for the benefit of all people, specifically related to Natural Language Processing (NLP)
- Build a structured Knowledge Base that combines Wikipedia and an extended named entity through a Research By Collaborative Contribution scheme

Research By Collaborative Contribution

__Dr Satoshi Sekine__ is the leader of the __SHINRA__ project that seeks to advance the use of knowledge within Natural Language Processing (NLP) tasks. He discusses his research in more detail, the importance of the resource collaborative contribution scheme and the next steps for his studies

Can you begin by telling us a little about the research you are currently engaged with?

I am the team leader within the Language Information Access Technology Team at the RIKEN Center for Advanced Intelligence Project in Japan. The project has many aims, but much of our work is centred on developing innovative technologies for the benefit of all humans, specifically those related to Natural Language Processing (NLP), which is a branch of Artificial Intelligence (AI).

What does NLP mean?

NLP is a way that machines can understand human language - which has many benefits, including enabling humans and machines to talk to each other and make sense of one another. One notable aspect of our work is concerned with creating a knowledge base that can function as an effective resource for NLP.

Can you explain in layperson terms what NLP involves and why it is such an important subject to know more about?

NLP can be thought of as a subfield that combines linguistics, computer science and AI. By giving machines the ability to understand text and spoken words in a similar way to humans, we can advance the field of computing to an unprecedented extent. Think about the ambiguities that exist between humans when they communicate - but facial expressions and tone can help with our understanding - we do not merely rely on what is said or written. If a computer can understand the unbelievable amount of information that exists on the Internet, then it will lead to efficiencies that are simply not possible if we just rely on humans.

Your main research topic is SHINRA. Can you talk a little more about what this involves?

SHINRA is a collaborative contribution scheme that utilises Automatic Knowledge Base Construction (AKBC) systems. This project aims to create a huge and well-structured Knowledge Base essential for many NLP applications, such as QA, dialogue systems and explainable NLP systems. We are aiming for a top-down design and a bottom-up population for this project, meaning we will use a name ontology that has been well designed and then populated this by crowdsourcing or using AI. SHINRA is centred on building a structured Knowledge Base that combines Wikipedia and an extended named entity through a Research By Collaborative Contribution scheme. Part of this is the SHINRA2020-ML task, which is to categorise 30-language Wikipedia pages into fine-grained named entity categories, called 'Extended Named Entity (ENE)'.

What are you hoping to focus on in the coming years?

SHINRA is our most pressing concern, but in the future we need to extend its scope to include knowledge of more general things. Then there is knowledge such as when it rains people use an umbrella. All of this knowledge needs to be machine readable - we need to think about what this means for machines because we have to think differently in terms of how machines understand as opposed to humans. ▶

Enabling machines to make sense of Wikipedia

*A team of researchers are working on the **RIKEN Center for Advanced Intelligence SHINRA** project which aims to achieve scientific breakthroughs and make meaningful contributions to the welfare of society and humanity using the development of innovative technologies*

Everyone knows about Wikipedia. Many of us have likely been chastised for using it in our essays and articles, not least because of its reputation as being rather unreliable. However, the fact is that Wikipedia is a beautiful idea - a community where people write pages of information about anything and everything, readily verify the information that is contained within those pages, ensure that claims can be referenced (and if they cannot, explain that a citation is needed), and basically provide an exhaustive resource that could never fully be pored over by any human, even if they were to live a thousand lifetimes.

This latter point is in stark contrast to computers, which can trawl through information at a rate that is many times faster than any human. Put simply, there is too much information 'out there' for any one individual to make sense of, but if we can somehow find a means of utilising machines, then the potential benefits are great. Unfortunately, the vast majority of information hosted on Wikipedia is written in such a way that makes it clear it is for people to read, rather than for a machine to utilise and manipulate.

Now a team of researchers have come together to create a project called SHINRA which is centred around a Resource by Collaborative Contribution (RbCC) scheme. The project aims to ensure that the information on Wikipedia can be structured in such a way that makes it possible for computers to use. The findings could have significant positive impacts on humanity, not least because of the development of an explainable AI system that uses deep learning and structured knowledge.

WORKING TOGETHER TO CONSTRUCT A RESOURCE

Dr Satoshi Sekine is the team leader of SHINRA, the project which focuses on AI, deep learning and Natural Language Processing (NLP). He is based within the Language Information Access Technology Team at the RIKEN Center for Advanced Intelligent in Japan. 'The ultimate aim of SHINRA is to build a structured Knowledge Base combining Wikipedia and an Extended Named Entity (ENE),' highlights Sekine. 'We hope to achieve our aim using a RbCC scheme, which is a novel endeavour for working together to construct a resource to advance the field of NLP and AI,' he says.

ENE is a hierarchy that was designed and developed to meet the increasing need for a wider range of named entity types. This hierarchy is divided into three main classes: name, time and numerical expressions. Through the team's observations, they have determined that any question on any specific matter generally fits within one of these three categories and by using such a hierarchy, much of what we consider to be common knowledge in newspaper articles, encyclopedia entries and Wikipedia pages can potentially be understood by machines engaged in deep learning.

The team worked tirelessly to define these classes based on a criterion of frequently occurring words and noun phrases, which were then categorised into a class according to the meaning and usage of each word and/or phrase. 'We extracted the candidate named entity expressions from newspaper, Wikipedias and many other types of documents,' outlines Sekine. 'More specifically, entities and proper nouns and numerical expressions were extracted from literally thousands of different contexts, before we assigned appropriate concept class labels,' he says. From there, the team worked to use references to existing online thesaurus entries and ontology sites to find information that matched the hierarchy.

Dr Satoshi Sekine; speaking at SHINRA project meeting

> *"We hope to achieve our aim using a RbCC scheme, which is a novel endeavour for working together to construct a resource to advance the field of NLP and AI"*

BUILDING A STRUCTURED WIKIPEDIA

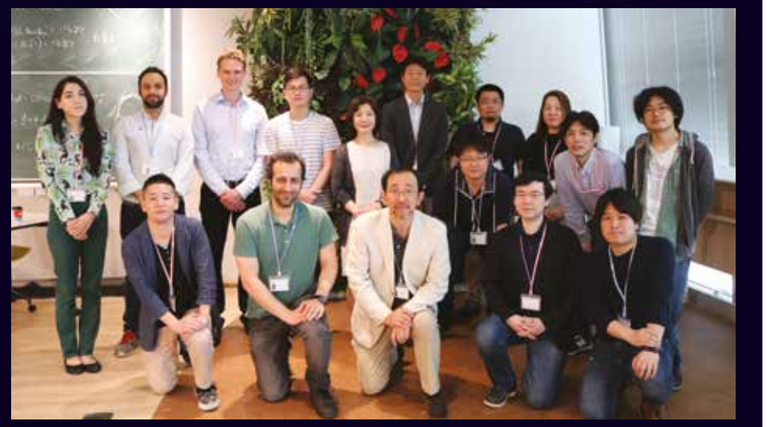
The ultimate goal of the project is building a structured Wikipedia. However, to achieve this, the team believes there are three tasks that need to be performed: categorisation, attribute extraction and attribute value linking. The first task is centred on categorising Wikipedia entities into ENE categories, which was originally done in Japanese using a machine learning method, with human experts checking the results. This was concluded in 2017, but a couple of years later the team updated it to reflect the 2019 version of Wikipedia. Ultimately, the scope has been expanded to reflect 30 languages, with the team working through the RbCC scheme on machine learning tasks.

For attribute extraction, the idea is to extract values of pre-defined attributes for each category from each Wikipedia entry. In 2018, the team achieved this for five categories, then 30 categories in 2019 and an additional 47 categories last year. Finally, for attribute value linking, the team is working on mapping the attribute values extracted from Wikipedia articles to the corresponding entities. This will be done for seven categories throughout 2021.

A VALUABLE KNOWLEDGE BASE

The SHINRA project commenced in 2017 and is still ongoing. One of the main ideas is to create a Knowledge Base scheme using the RbCC approach. Sekine and the rest of his team are acutely aware of how popular and ever-present AI will become in the near future - and there are already many examples of it permeating our daily lives, whether that be at work or in our leisure time. 'One of the most crucial aspects of the research the team is undertaking is to ensure that the technology employed in everyday scenarios is capable of explaining its processes and/or decision making and that this is based on the RbCC approach,' explains Sekine. 'If this can be achieved then we can truly begin to utilise all of the information that is out there,' he says.

In many ways Sekine and the team have only just begun what they are ultimately hoping to achieve. The key idea is that they create a Knowledge Base that can be used by them in future studies (or even different researchers in different studies) to advance the ability of computers to not only make sense of what it is they are reading, but explain the rationale of their thought. It may sound scary, but as the capacity of machine learning improves, the simple truth is that humans can learn from another sentient 'thing' for the first time in history. ●



Members of the Language Information Access Technology Team (including intern students from around the world)

Project Insights

PROJECT MEMBERS

Satoshi Sekine (Chair), Kouta Nakayama, Akio Kobayashi, Masako Nomoto, Maya Ando, Michiko Goto, Asuka Sumida, Koji Matsuda, Yu Usami, Masatoshi Suzuki, Yukino Baba and Akiva Miura

We appreciate all contributions from the participants of SHINRA project

CONTACT

Dr Satoshi Sekine

E: satoshi.sekine@riken.jp
W: http://shinra-project.info/shinra2020ml/?lang=en