2. 【現在までの研究状況】(図表を含めてもよいので、わかりやすく記述してください。様式の変更・追加は不可(以下同様))
① これまでの研究の背景、問題点、解決方策、研究目的、研究方法、特色と独創的な点について当該分野の重要文献を挙げて記述してください。
② 申請者のこれまでの研究経過及び得られた結果について、問題点を含め①で記載したことと関連づけて説明してください。
なお、これまでの研究結果を論文あるいは学会等で発表している場合には、申請者が担当した部分を明らかにして、それらの内容を記述してください。

【研究の背景】

With the increase in cultural communication and trade between countries, the demand for translation is also increasing. Compared with translating manually, machine translation (MT) can generate good-quality translations quickly and at low cost.

Educational domain MT is important because most open online courses are in English. We need to add subtitles of other languages so that people speaking other languages can benefit from high-quality educational resources.

Current data-driven MT [1] can generate high-quality translations for domains with a lot of data, such as news. But there is no high-quality corpus in the educational domain. Without enough data, we cannot construct (train) a MT model of satisfactory performance. We proposed an educational corpus and a method to use the corpus to train a high-quality MT model.

【問題点】

1. No corpus. There is no high-quality corpus in the educational domain.
2. No suitable training method. We have multiple corpora. If we train the MT model using the mixture of all the data, the model will give an average performance among all domains. It is not the optimal performance in educational domain.

【研究目的】

To construct a corpus and design a training method for educational domain MT.

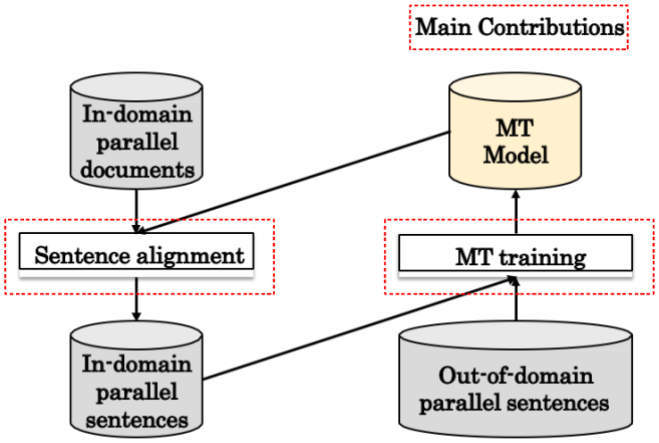

Figure 1: Overview framework of proposed methods.

【解決方策・研究方法】

Figure 1 shows our framework, it contains:

1. Corpus construction. Website Coursera is providing free lectures. We crawl and clean the data from it. For each lecture, there are parallel subtitle documents of multiple languages. We extract parallel sentences from parallel documents. Finally, we divide the corpus into train/validation/test set.
2. Better training method. We have several larger out-of-domain corpora and a smaller educational domain corpus. To make a MT model optimal for educational domain, in the former stages, we use out-of-domain corpora. In the latter stages, we use educational domain corpus to train the model so that it is optimal for the educational domain.

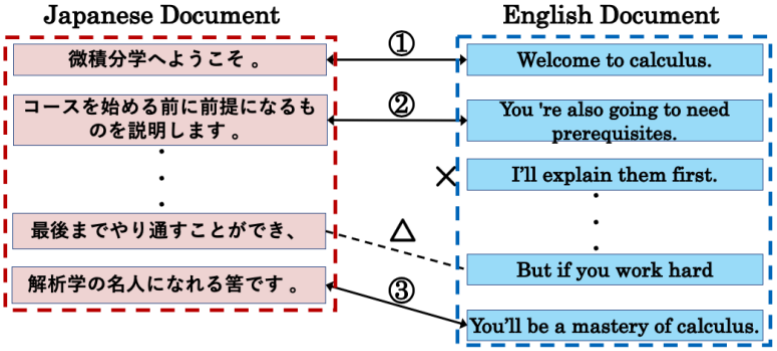

Figure 2: Three parallel sentences from parallel documents.

【特色と独創的な点】

1. Sentence alignment algorithm (Figure 2) using semantic information. The algorithm is to mine parallel sentences from parallel documents. It makes use of the meaning of words rather than the surface n-gram information. It can improve the sentence alignment accuracy thus improve the quality of corpus. It can also apply to other corpus construction tasks.
2. Multiple stages training method. Compared with previous methods [2], our training method made the model optimal for educational domain while preserving general knowledge as much as possible. The model learns general information in the former stages and become optimal for educational domain in the latter training stages.

申請者登録名　　ソウ　カイエツ

DC

（現在までの研究状況の続き）

【申請者のこれまでの研究経過及び得られた結果】

1 Japanese-English educational domain corpus. It contains more than 40,000 parallel line pairs in the training set, 2000 and 500 manually checked sentence pairs in the test and dev set（発表文献[4]). With this corpus, we saw about 2 BLEU score improvement for the educational domain MT.

2 Several training methods for educational domain MT.

2.1 A multistage fine-tuning method. It is useful when there is more than one corpus. It showed 2-5 BLEU score improvement compared with previous training methods (発表文献[4]). For low-resource domains, such as educational domain, we obtained better translation quality. As the first author, I've done most part of the work.

2.2 A method leveraging data in other languages. Experiments showed maximum 5-8 BLEU score improvement in low-resource situation for Japanese-English translation (発表文献[6]). This training method can also be applied in educational domain MT. As the first author, I've done most part of the work.

2.3 A pre-training method that involves Japanese linguistic information. It showed maximum 7 BLEU score improvement for language pairs involving Japanese (発表文献[5]). It is especially useful for domain with few data such as educational domain. I supported the experiments and paper writing in this work.

【参考文献】

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS, pages 6000–6010.
[2] Miceli Barone, A. V., Haddow, B., Germann, U., and Sennrich, R. (2017). Regularization Techniques for Finetuning in Neural Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1489–1494, Copenhagen, Denmark, September. Association for Computational Linguistics.

3. 【これからの研究計画】

(1) 研究の背景

これからの研究計画の背景、問題点、解決すべき点、着想に至った経緯等について参考文献を挙げて記入してください。

【研究の背景・着想に至った経緯】

In recent years, MT needs in many low-resource languages are increasing rapidly. To make more people accessible to different cultures and knowledges, it is necessary to construct a multilingual MT model for low-resource languages.

We need lots of parallel data to build a high-quality MT model. We have large corpora for Japanese-English, but no such corpora for most other language pairs.

We propose a method to construct multilingual corpora. To maximize the use of the corpus, we propose a set of low-resource training techniques. The current training methods [3] only use data from one language pair. We found data from assisting languages can help the training. The framework is shown in Figure 3.

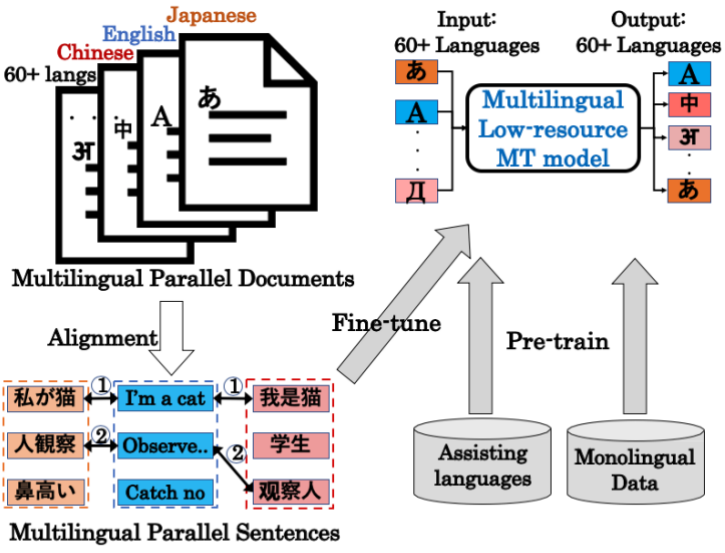

Figure 3: Multilingual corpus construction and domain adaptation for multilingual low-resource MT model construction.

【問題点】 No multilingual corpus. And the training requires lots of parallel data.

【解決すべき点】

To construct a multilingual corpus, we need to collect high-quality raw data and alignment sentences in different languages in high accuracy. We also need training method to leverage data from assisting languages, which is easier to get.

申請者登録名　　ソウ　カイエツ

DC

(2) 研究目的・内容（図表を含めてもよいので、わかりやすく記述してください。）

① 研究目的、研究方法、研究内容について記述してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、具体的に記入してください。
③ 所属研究室の研究との関連において、申請者が担当する部分を明らかにしてください。
④ 研究計画の期間中に異なった研究機関（外国の研究機関等を含む。）において研究に従事することを予定している場合はその旨を記載してください。

【研究目的】

To provide a general method for multilingual corpus construction and design a MT training method for low-resource languages.

【研究内容・方法】

1. Multilingual corpus construction. We first collect large-scale data of 60+ low-resource languages. Then we use multilingual models [4] to measure the similarity of sentences from different languages. Finally, we use dynamic programming algorithm to extract the global best sentence match with quality insurance.

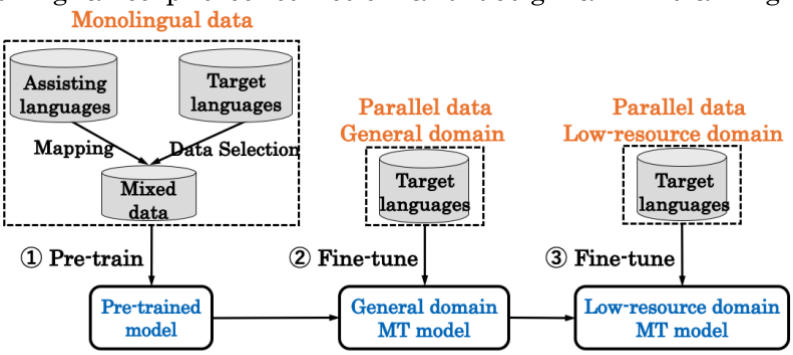

Figure 4: Training a low-resource MT model.

2. Training methods for low-resource languages. As shown in Figure 4, we use data from assisting languages and improve cognate sharing by mapping. As shown in Figure 5, Chinese can act as assisting language of Japanese and we can generate synthetic Japanese data through a mapping table [5]. We also apply data-selection in the pre-training process. We then design a multistage fine-tuning process to train a low-resource MT model optimal for a certain domain.

Figure 5: Assistant language mapping

【参考文献】

[3] Rico Sennrich and Biao Zhang. Revisiting low resource neural machine translation: A case study. In Proc. of ACL 2019, pages 211– 221. [4] Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. arXiv:1901.07291. [5] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese. In Proc. LREC'12, pages 2149–2152.

(3) 研究の特色・独創的な点

次の項目について記載してください。
① これまでの先行研究等があれば、それらと比較して、本研究の特色、着眼点、独創的な点
② 国内外の関連する研究の中での当該研究の位置づけ、意義
③ 本研究が完成したとき予想されるインパクト及び将来の見通し

【本研究の特色、着眼点、独創的な点】

Both the multilingual corpus construction method and the training method for low-resource language are novel. Combining them, we can greatly advance low-resource machine translation.

1 Multilingual corpus construction.

1.1 Multilingual corpus containing data of more than 60 languages. Most of the previous corpora [6] focus on bilingual corpus construction. We focus on multilingual corpus construction. We have data of more than 60 languages, where most of them are low-resource languages. Our propose method can automatically construct bilingual parallel corpus of any language pairs.

1.2 Robust alignment algorithm for multilingual sentences alignment. Previous alignment algorithms [7] focus on surface n-graph information. There are low-frequency words in data of low-resource languages. Our model turns sentences in different languages into vectors in the same semantic space. We can measure the similarity of two sentences by measuring the distance of two vectors. Our alignment method is more robust and can generate high-quality sentence pairs.

2 Assisting languages in the model training process. This will be the first research exploring the importance of cognate sharing in the pre-training process. We found there are cognate sharing among several language pairs, such as Chinese-Japanese and Hindi-Urdu. We map the data of assisting language into interested language and generate synthetic data. This is especially useful in extreme low-resource situation.

（研究の特色・独創的な点の続き ）

【関連する研究の中での当該研究の位置づけ、意義】

1 We propose a method to automatically construct multilingual corpora. This general method can apply to any multilingual parallel documents. With our method, building multilingual corpora will become much easier. With larger multilingual corpora, low-resource MT will advance greatly.

2 The alignment algorithm is both an application and verification of the current multilingual pre-training techniques. It will motivate further research of multilingual models.

3 Our training method using data from assistant language will be the very initial research to explore how to leverage data from assisting languages in low-resource MT field.

【本研究が完成したとき予想されるインパクト及び将来の見通し】

For people speaking resource-rich languages such as English, there is much data thus high-performance MT models. Benefiting from the MT model, translation becomes easier and there is more data. But for people speaking low-resource languages, they are the people who need MT techniques most to reach the outer world. But unfortunately, there is less data in their languages, making it hard to build a high-quality MT model. It is a negative cycle.

Our research is exactly aiming to break the negative cycle. With our multilingual corpus and domain adaptation methods. We can build a better MT model of more than 60 languages. It can apply to educational domain so that people can explore online courses from top universities regardless of their geographical location. We also develop a MT system to add subtitles to lectures in Kyoto University. It can help foreigner students to understand Japanese lectures better.

【参考文献】

[6] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian Scientific Paper Excerpt Corpus. In Proc. of LREC'16, pages 2204–2208.

[7] Sennrich, R. and Volk, M. (2011). Iterative, MT-based Sentence Alignment of Parallel Texts. In Proc. of NODALIDA'11, pages 175–182.

(4) 研究計画

申請時点から採用までの準備状況を踏まえ、研究計画について記載してください。

	First year	Second year	Third year
Multilingual corpus construction	Collect large-scale raw data	Alignment algorithm development (Conference paper)	Construct multilingual parallel corpora (Journal paper)
Training methods for low-resource situation	Pre-training method leveraging data from assisting languages	Low-resource fine-tuning algorithm	Combine pre-training method and fine-tuning method (Journal paper)
MT system development	(Conference paper)	(Conference paper)	Build a low-resource MT model using multilingual corpora and the training method.

（一年目）

1. Raw data collection and data cleaning. We have collected multilingual data from Coursera website. There are totally 363K subtitle files from more than 4000 courses in more than 60 languages. There are more than 9K parallel English-Japanese documents and 2K Japanese-English-Chinese trilingual parallel documents. We are now planning to collect more subtitle data from other websites such as MOOC and YouTube.
2. Pre-training for low-resource situation. We will try using data from different assisting languages and find the relationship between cognate sharing and translation quality improvement.

（二年目）

1. Alignment algorithm for multilingual sentences. We plan to try several multilingual models and verify them through experiments. We will choose the model that gives the highest accuracy in multiple language pairs.
2. Low-resource fine-tuning method. We will propose several domain adaptation methods which fit the low-resource situation better.

（研究計画の続き ）

（三年目）

1. Corpus construction. We apply the alignment algorithm to the large-scale raw data. We will have bilingual parallel corpus for all language pairs and construct several small trilingual parallel corpora as test set. We will also public the document-level parallel data for document-level MT research.
2. Training methods combination. Combining the pre-training and multistage fine-tuning methods, the training method will improve low-resource MT quality greatly. And it can also apply to other low-resource situations.
3. MT system integration. We will apply the low-resource situation training method to the large-scale multilingual corpus. The final MT model can achieve practical level translation quality for multiple language pairs.

(5) 人権の保護及び法令等の遵守への対応

本欄には、研究計画を遂行するにあたって、相手方の同意・協力を必要とする研究、個人情報の取り扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究など法令等に基づく手続が必要な研究が含まれている場合に、どのような対策と措置を講じるのか記述してください。**例えば、個人情報を伴うアンケート調査・インタビュー調査、国内外の文化遺産の調査等、提供を受けた試料の使用、侵襲性を伴う研究、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験など**、研究機関内外の情報委員会や倫理委員会等における承認手続が必要となる調査・研究・実験などが対象となりますので手続の状況も具体的に記述してください。

なお、該当しない場合には、その旨記述してください。

該当なし

申請者登録名　　ソウ　カイエツ

ＤＣ

4.【研究遂行能力】研究を遂行する能力について、これまでの研究活動をふまえて述べてください。これまでの研究活動については、網羅的に記載するのではなく、研究課題の実行可能性を説明する上で、その根拠となる文献等の主要なものを適宜引用して述べてください。本項目の作成に当たっては、当該文献等を同定するに十分な情報を記載してください。
具体的には、以下(1)～(6)に留意してください。

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書（査読の有無を明らかにしてください。**査読のある場合、採録決定済のものに限ります。**）
著者、題名、掲載誌名、発行所、巻号、
pp 開始頁－最終頁、発行年を記載してください。

(2) 学術雑誌等又は商業誌における解説、総説

(3) 国際会議における発表（口頭・ポスターの別、査読の有無を明らかにしてください。）
著者、題名、発表した学会名、論文等の番号、場所、月・年を記載してください。（**発表予定のものは除く。**ただし、発表申し込みが受理されたものは記載してもよい。）

(4) 国内学会・シンポジウム等における発表
(3)と同様に記載してください。

(5) 特許等（申請中、公開中、取得を明らかにしてください。ただし、申請中のもので詳細を記述できない場合は概要のみの記載してください。）

(6) その他（受賞歴等）

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書

〈査読有り〉

[1] Li Jiang, Zhuoran Song, Haiyue Song, Chengwen Xu, Qiang Xu, Naifeng Jing, Weifeng Zhang, Xiaoyao Liang. Energy-Efficient and Quality-Assured Approximate Computing Framework Using a Co-Training Method, ACM Transactions on Design Automation of Electronic Systems (TODAES 2019), pp.59:1-59:25, (2019.11)

(3) 国際会議における発表

〈査読有り・ポスター発表〉（○発表者）

[2] Haiyue Song, Xiang Song, ○Tianjian Li, Hao Dong, Naifeng Jing, Xiaoyao Liang, Li Jiang. A FPGA Friendly Approximate Computing Framework with Hybrid Neural Networks (Abstract Only), In Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2018), pp.286, Monterey, CA, USA, (2018.2). 交換留学中のため発表は共著者

〈査読有り・口頭発表〉（COVID-19 の影響で[4][5]の発表は延期）

[3] Haiyue Song, Chengwen Xu, Qiang Xu, ○Zhuoran Song, Naifeng Jing, Xiaoyao Liang, Li Jiang. Invocation-driven neural approximate computing with a multiclass-classifier and multiple approximators, In Proc. of the International Conference on Computer-Aided Design (ICCAD 2018), pp.50, San Diego, CA, USA, (2018.11). Acceptance rate = 25%. 卒業したため発表は共著者

[4] ○Haiyue Song, Raj Dabre, Atsushi Fujita and Sadao Kurohashi. Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation, Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp.3640-3649, Marseille, France, (2020.5).

[5] ○Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song and Sadao Kurohashi. JASS: Japanese-specific Sequence to Sequence Pre-training for Neural Machine Translation, Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp.3683-3691, Marseille, France, (2020.5).

(4) 国内学会・シンポジウム等における発表

<査読なし・口頭発表>

２報、筆頭１報と共著１報、いずれも発表者

【発表(印刷)前】

(1) 国際会議に採録決定されたもの

<査読有り・口頭発表>

[6] Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi and Eiichiro Sumita. Pre-training via Leveraging Assisting Languages for Neural Machine Translation, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL 2020 SRW), pp., Seattle, Washington, United States, (2020, 7). Accepted papers list: https://sites.google.com/view/acl20studentresearchworkshop/accepted-papers

5.【研究者を志望する動機、目指す研究者像、アピールポイント等】

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、研究者を志望する動機、目指す研究者像、アピールポイント等を記入してください。

【研究職を希望する動機】

Research can satisfy my curiosity and I hope my research can help people worldwide. I'm always eager to explore new things in different fields. I choose low-resource MT as my research topic because it can help people whose native language is not English to access high-quality resources. I believe that the research post in Japan is the best and the only choice for me.

【目指す研究者像】

1. I would like to keep research as both interest and duty. I started learning computer science as interest since high school and enjoyed research from undergraduate period. It's also a duty because I found that many new technologies have not touched people who need them most, such as people living in undeveloped areas. I hope my research can help them.
2. I will try my best to turn research results into positive impact in the world. During internship in companies, I found there is a large gap between research and practice. Bridging the gap between research and real-world impact is always an important task for me.
3. I aim to become a researcher to promote communication and understanding among different cultures. I have made many foreigner friends and gradually understand the importance of intercultural communication. I hope my MT research can make the communication easier.

【自分の長所】

1. Positive. I have good hope for the future and eager to create it actively. I can recover quickly from failure, such as preparing more experiments actively after our paper got rejected.
2. High productivity. I have 2 publications on top conferences and 1 paper on top journal during the undergraduate period. In the last 6 months, 5 of my papers got accepted (three of them are peer reviewed) by international and domestic conferences, and I am still preparing two journal papers now. With more experience, I can be more productive during my PhD study.
3. International cooperation ability. I always collaborate with researchers from different countries. One of the publication (発表文献[5]) involves the collaboration of researchers from Japan, India, French, and China. And I can use Japanese, English, and Chinese to communicate with them smoothly. During social gathering of conferences, I always try to make friends with other researchers.
4. Board research interests. I have experience and publications (reports) in NLP, CV and DNN acceleration fields. Ideas from different fields mix in my mind and brilliant ideas born.

【自己評価する上で、特に重要と思われる事項】

1. Many publications. Recently, three of our papers get accepted to international conferences. They are about building a dataset to improve Japanese-English educational domain machine translation (発表文献[4]), using Japanese linguistic information to improve machine translation (発表文献[5]), and using Chinese data to improve Japanese MT (発表文献[6]).
2. Started doing research from undergraduate period. During the third and fourth year in college, I went to the lab almost every day and have several publications (発表文献[1-3]).
3. I took internships in National Institute of Information and Communications Technology (NICT) and LINE. I produced a paper (発表文献[4]) in the two-month's internship in NICT.
4. I learned algorithms and data structures by myself in high school and got silver medal in China's National Olympiad in Informatics 2013 (Rank: 78/280).
5. I started learning Japanese from 2nd year in the university. In the daytime I attend computer science classes and at night I attend the Japanese classes. During the exchange program at Nagoya university, I passed N1 level of Japanese Language Proficiency Test (JLPT). I attended the international course during my bachelor and master's study, where most classes are in English. I got 97/120 scores in Test of English as a Foreign Language (TOEFL) exam.
6. Research as daily life. I like reading papers in the bath and on the train. I read about 50 papers on the train during the internship in NICT, which turned to material of my own paper.

Personal Homepage: https://shyyhs.github.io

申請者登録名　　ソウ　カイエツ