Sentiment Analysis with Language Models

Kazutaka Shimada
Department of Artificial Intelligence
Kyushu Institute of Technology

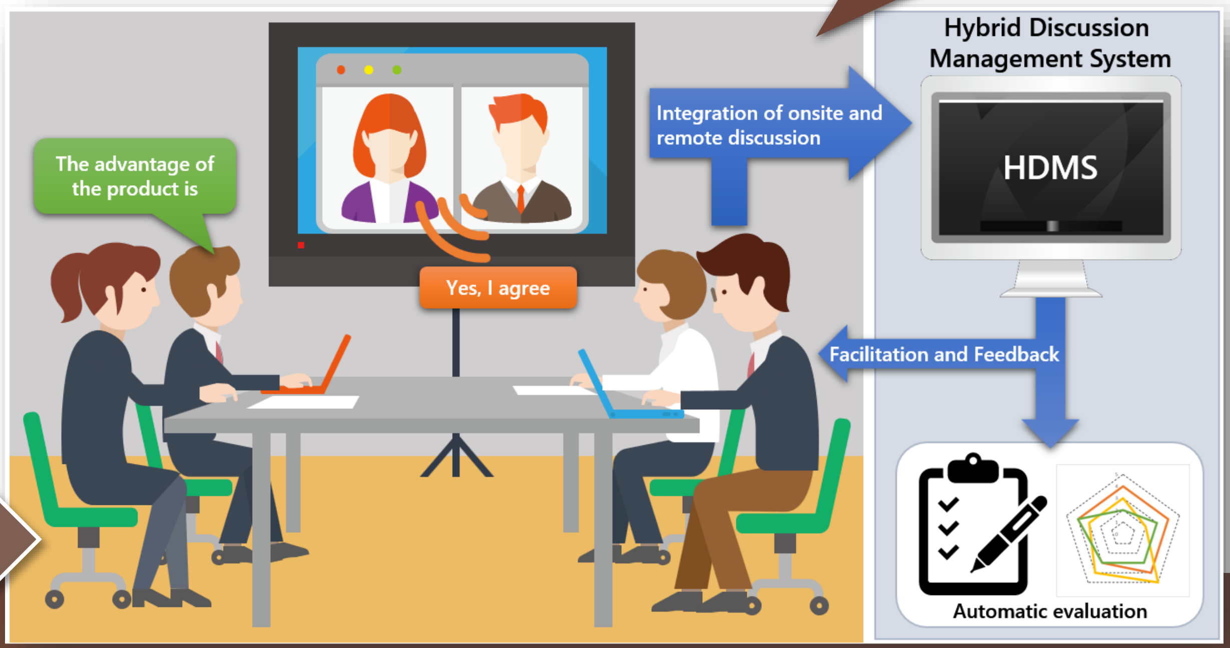
THE 4TH
INTERNATIONAL CONFERENCE
ICICYTA
2024

Self-introduction: Kazutaka Shimada

- Affiliation: Department of Artificial Intelligence, Kyushu Institute of Technology
- Research topic: Natural Language Processing (NLP)
  - Text analysis, information extraction, and summarization

Multiparty conversation understanding

Current Interest



Today's contents

- Main topic
  - Sentiment analysis
    - Sentence classification into
      - Positive/negative and 1-5 stars (seeing stars)
      - Sarcastic/non-sarcastic
      - Offensive/non-offensive, and so on
    - Word extraction, such as aspect/target of sentiment, from sentences
  - Additionally, some related tasks about data on the web
- First of all, what's a (large) language model?
  - Basic idea and current approaches in the neural era
- Then, sentiment analysis with NLP techniques

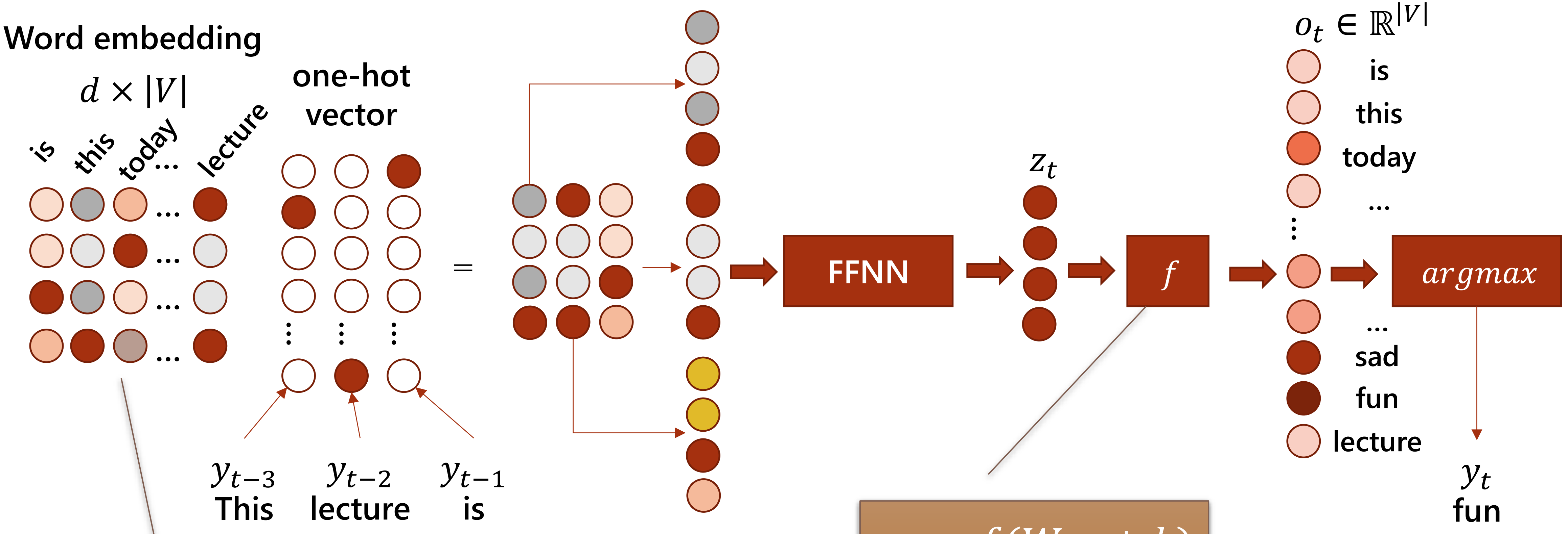
What is a language model?

Sentiment Analysis with Language Models

Language model

- Language model
  - Assigning a probability for the likelihood of a given word to follow a sequence of words
    - John eats an ______
    - $P(w_i|w_1, w_2, \ldots, w_{i-1})$
- Traditional approach
  - For a sentence with k words and the bi-gram model
    - $P(w_1 \ldots w_k) = \prod_{i=1}^{K} P(w_i|w_{i-1})$
    - bi-gram: for-a, a-sentence, sentence-with, …

Feed Forward Neutral Network: FFNN

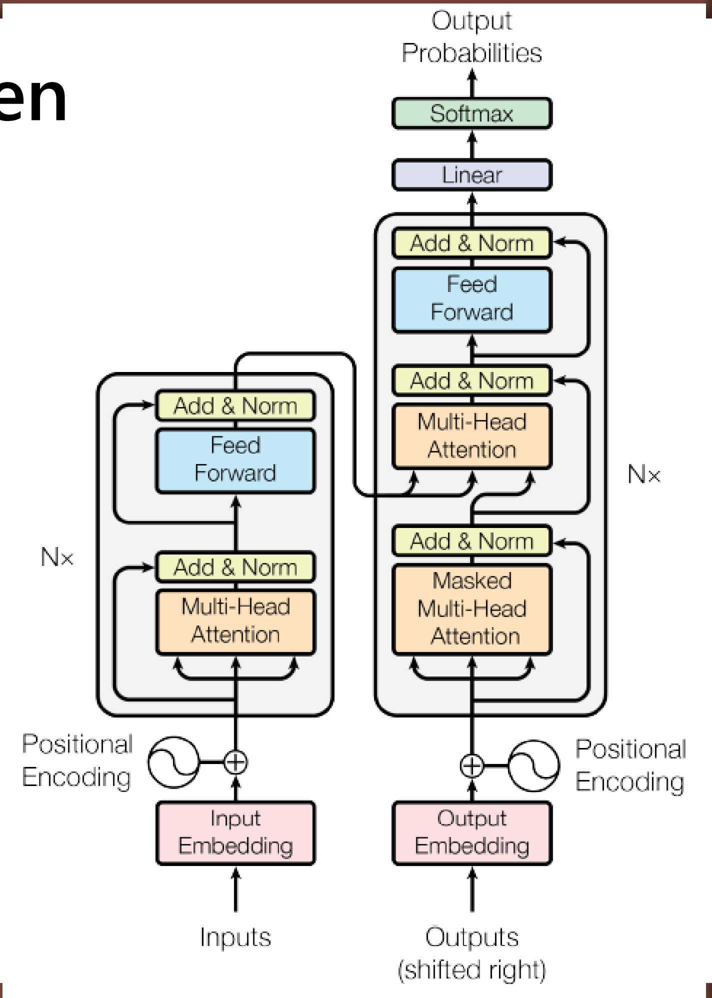


This example: 4 dims

$$o_t = f(Wz_t + b)$$
$$W \in \mathbb{R}^{|V| \times d}$$
$$b \in \mathbb{R}^{|V|}$$

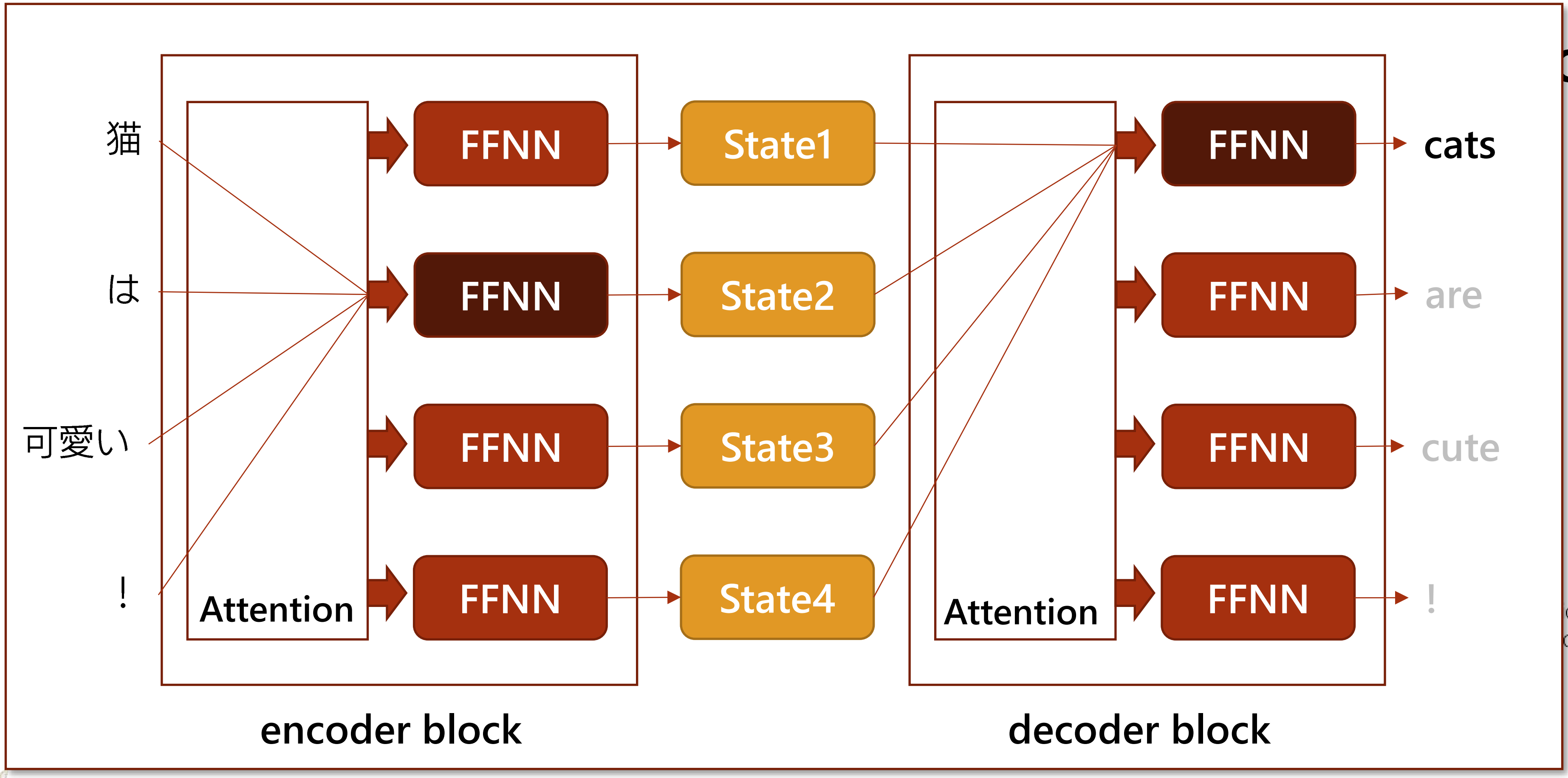
Transformers: Attention is All You need

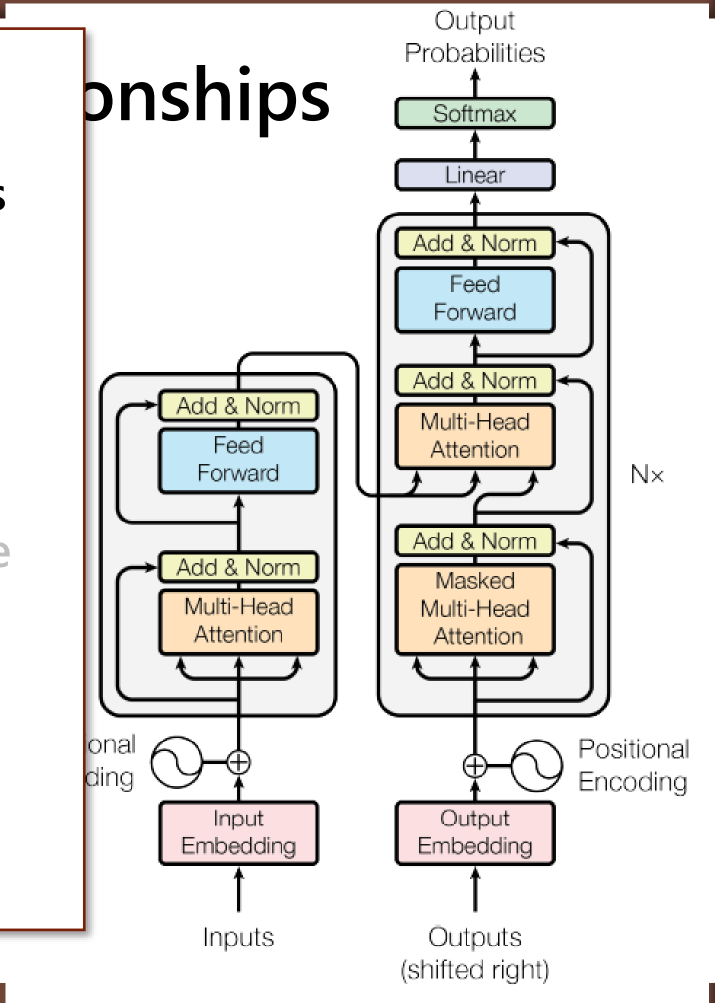
- It is difficult to capture the relationships between distantly located words by sequential models such as LSTM and RNN
- Moreover, they are not suitable parallelization
- Transformer
    - Use the attention mechanism only
    - Fundamental model for current LLMs
        - Query-Key-Value
        - Position encoding
        - Self / Multi-head attention



This figure is from the original paper

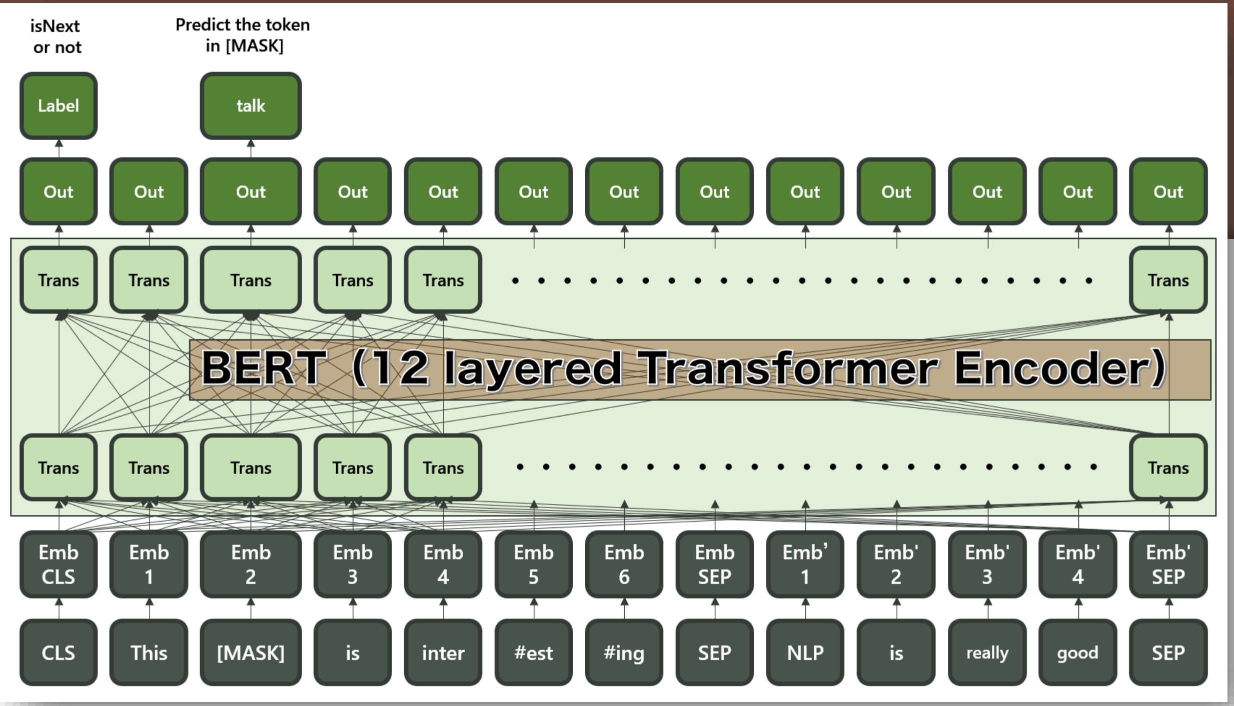
Transformers: Attention is All You need





This figure is from the original paper

BERT: pre-training



- Masked LM
  - 80%: [mask], 10%: token, 10%: unchanged
  - `my cat is cute`

```
→ my cat is [MASK]
→ my cat is apple
→ my cat is hairy
```

- Next sentence prediction: IsNext or NotNext

```
Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
  Label = IsNext
Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds
[SEP]
  Label = NotNext
```

Fine-tuning

This figure is from the original paper

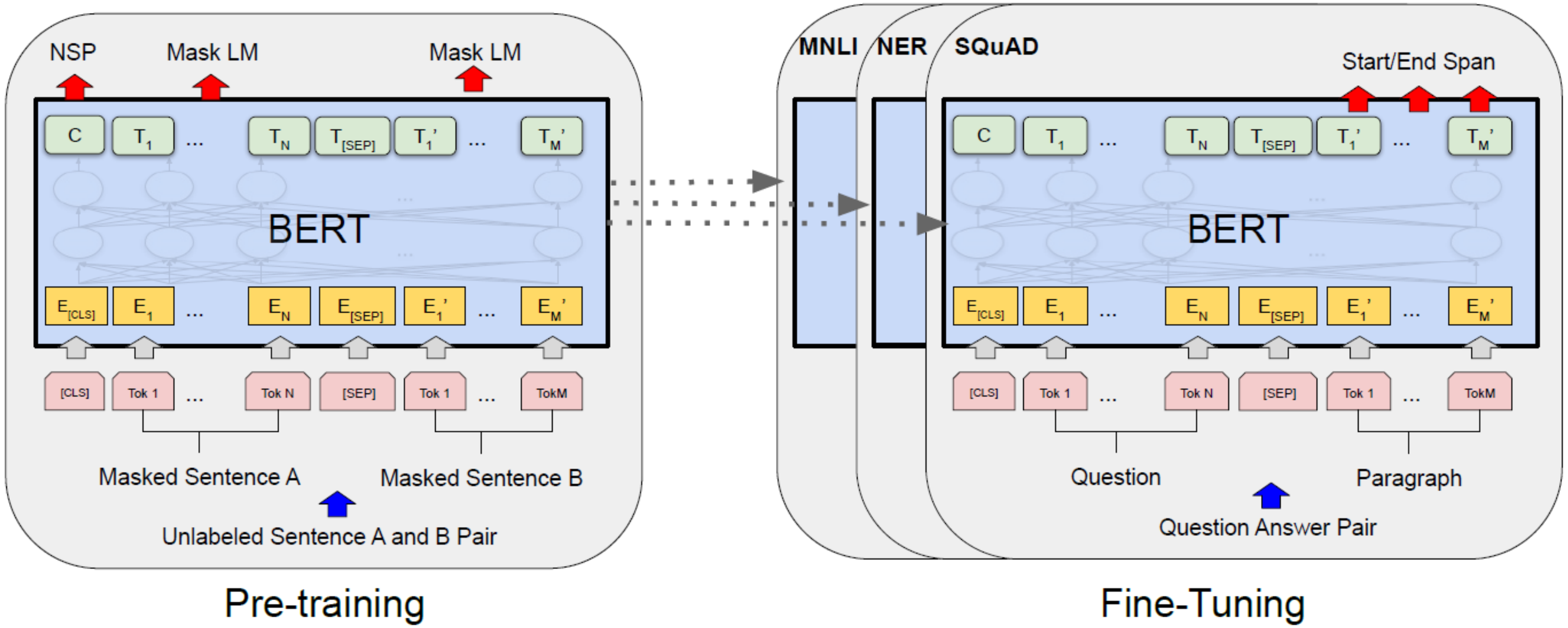
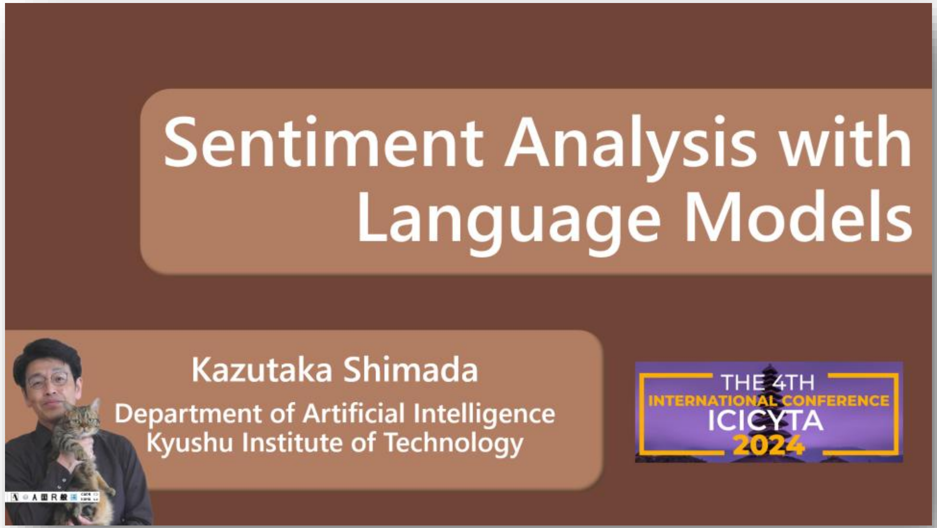


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Basic Model and Ideas

- Basic model in this talk is BERT
  - BERT is a masked language model, but Large language model (LLM)???
    - # of parameters: approximately 0.1 billion .... even GPT-3 is 175B
  - The title of this talk is "sentiment analysis with language models"
- Basic ideas in my lab (this talk)
  - Utilize something for BERT
  - Apply something to BERT
  - and sometimes using LLM

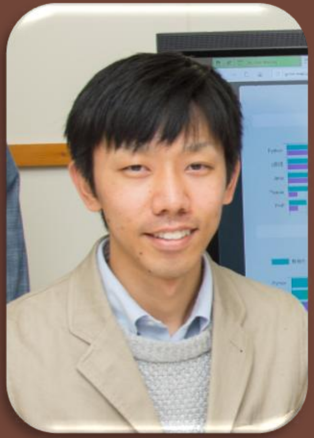


The models are just a tool, and they aren't purpose

Today's contents

- Main topic
  - Sentiment analysis
    - Sentence classification into
      - Positive/negative and 1-5 stars (seeing stars)
      - Sarcastic/non-sarcastic
      - Offensive/non-offensive, and so on
    - Word extraction, such as aspect/target of sentiment, from sentences
  - Additionally, some related tasks about data on the web
- First of all, what's a (large) language model?
  - Basic idea and current approaches in the neural era
- Then, sentiment analysis with NLP techniques

Satoshi Hiai (PhD candidate 2018-2020)



Sarcasm detection using relation information

Sentiment Analysis with Language Models

Sarcasm detection

- Sarcasm is one of sentiment expressions
  - Very rare but important for understanding the language
  - Usually context-dependent

He is a good boss who gives his staff homework to do on the weekend.

  - Often said "sarcasm contains not only positive expressions but also negative expressions
- We focus on parallel relations, such as "boss" and "staff"
  - "professor" and "student", "parent" and "child"

Bi-LSTM with relation vectors

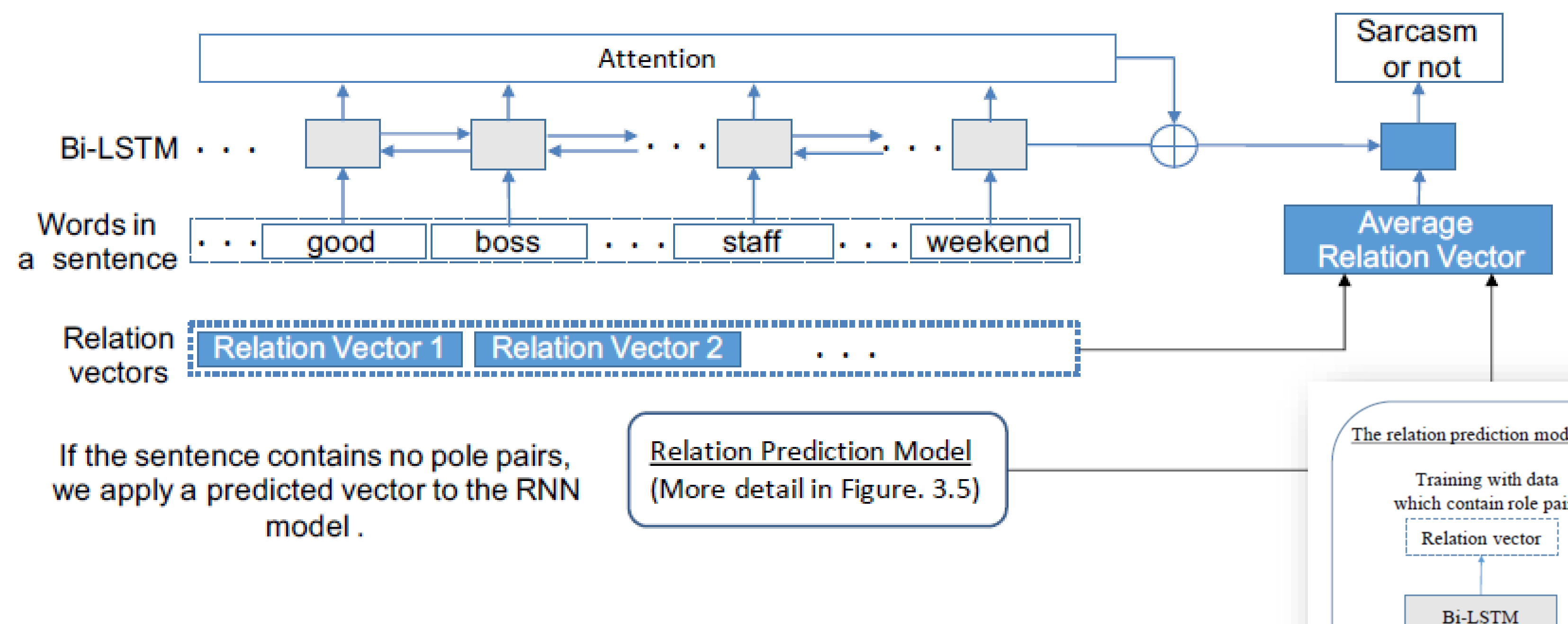


Figure 3.4 Detail of the RNN model with relation vector.

Model	F1-score
SVM	0.751
Bi-LSTM with word embeddings	0.772
Ours (with relation vectors)	0.802

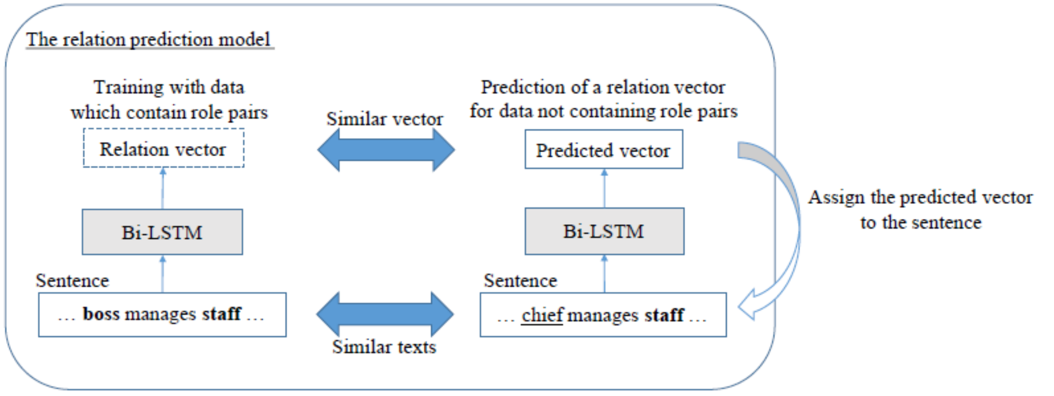


Figure 3.5 Overview of relation prediction model

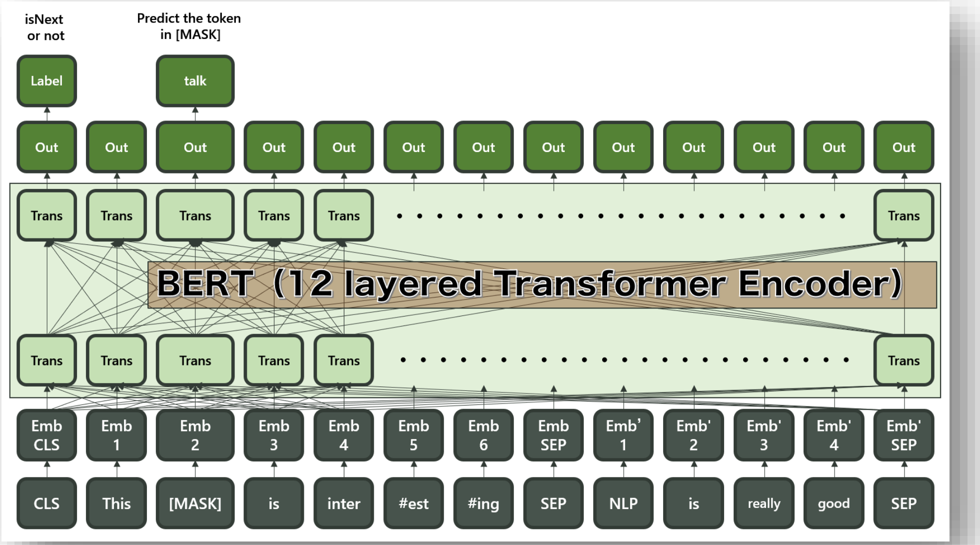
Li Zhenming (PhD candidate 2023-now)

Offensive language detection with knowledge extension

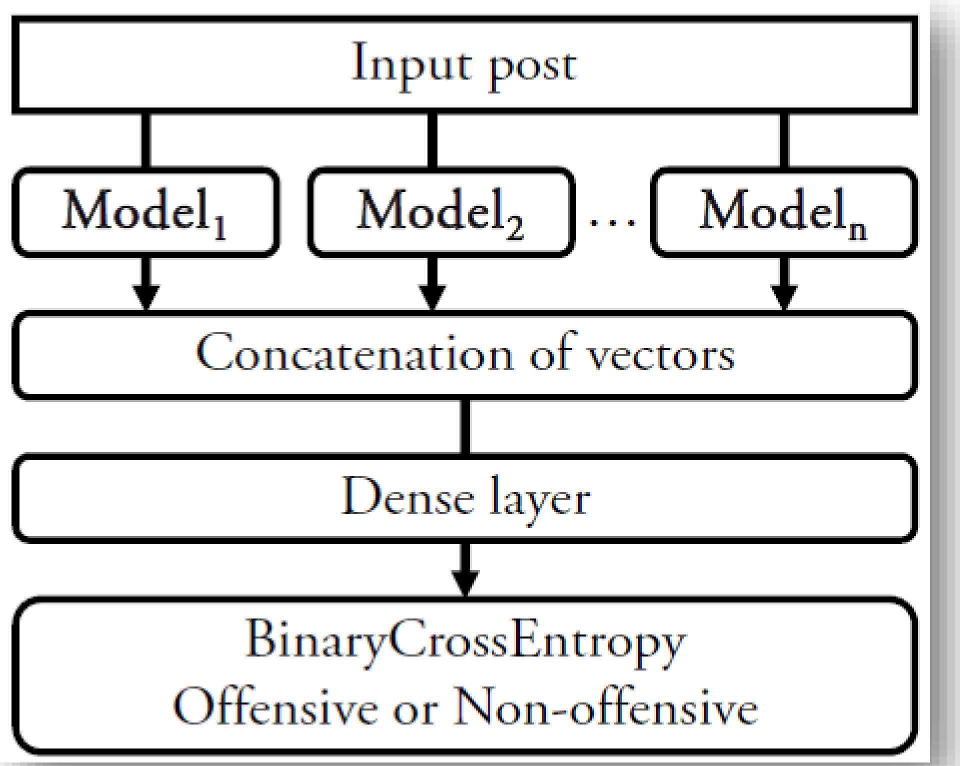
Sentiment Analysis with Language Models

Offensive language detection

- Offensive posts/comments on Social Media bring mental damage to victims
  - As a classification task by BERT
- Combining several models
  - BERT, DeepMoji, and HateBERT
- Combining several datasets
  - Extracting similar instances from another dataset
  - Utilizing them to fine-tuning of the target task

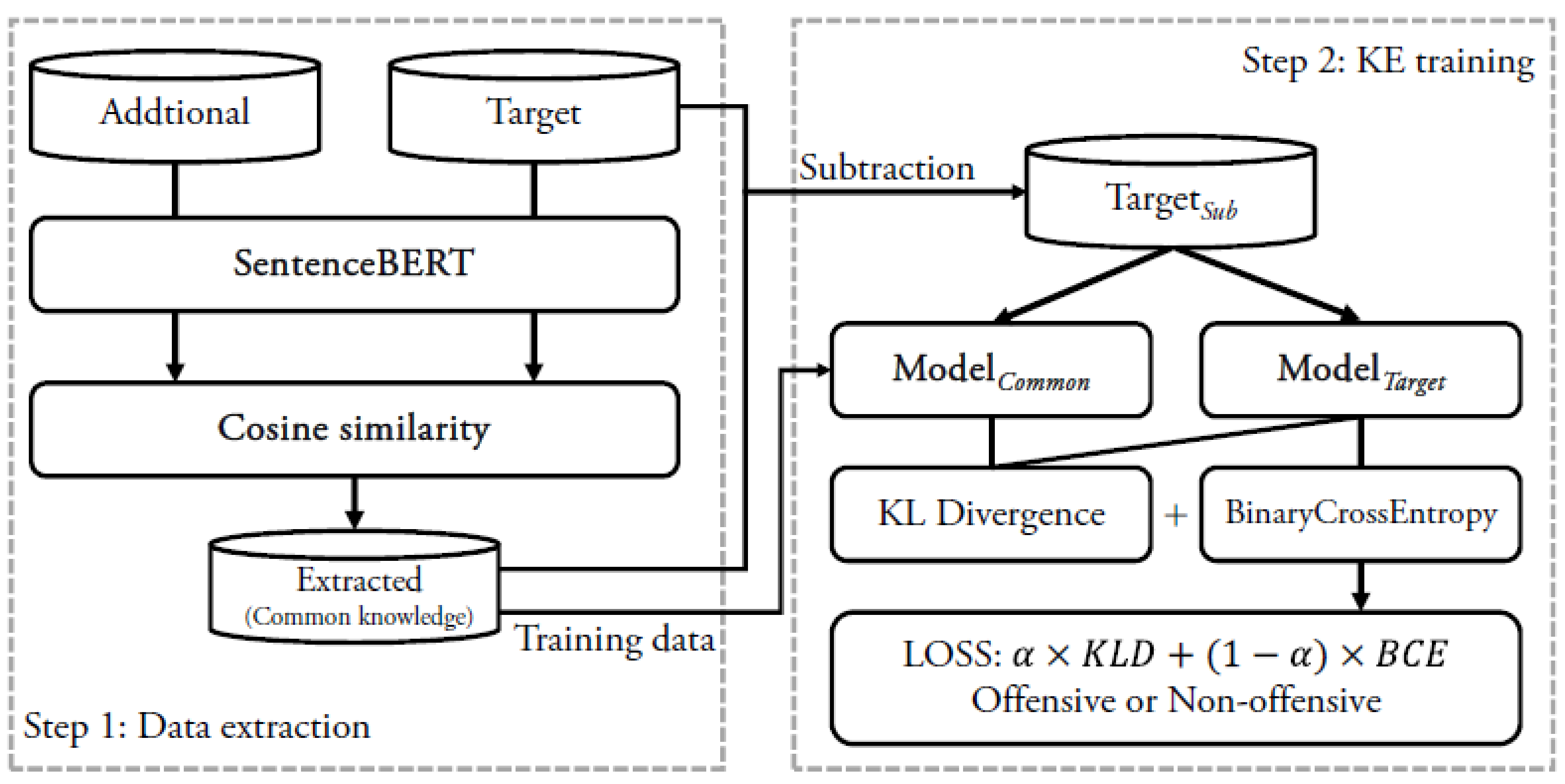


Model combination



Model	OLID	Curious Cat	AskFM
DeepMoji(D)	0.659	0.638	0.589
BERT(B)	0.676	0.489	0.543
HateBERT(H)	0.706	0.701	0.598
D+B	0.654	0.488	0.530
D+H	0.697	0.716	0.603
B+H	0.723	0.667	0.590
D+B+H	0.702	0.726	0.610

Knowledge extension



Model	Curious Cat Only	KE with OLID	KE with AskFM
DeepM	0.638	0.604	0.475
BERT	0.489	0.522	0.537
HateB	0.701	0.535	0.724

Note that the best score of the Curious Cat dataset is 0.726 by D+B+H in the previous slide

Niraj Pahari (PhD candidate 2022-2024)

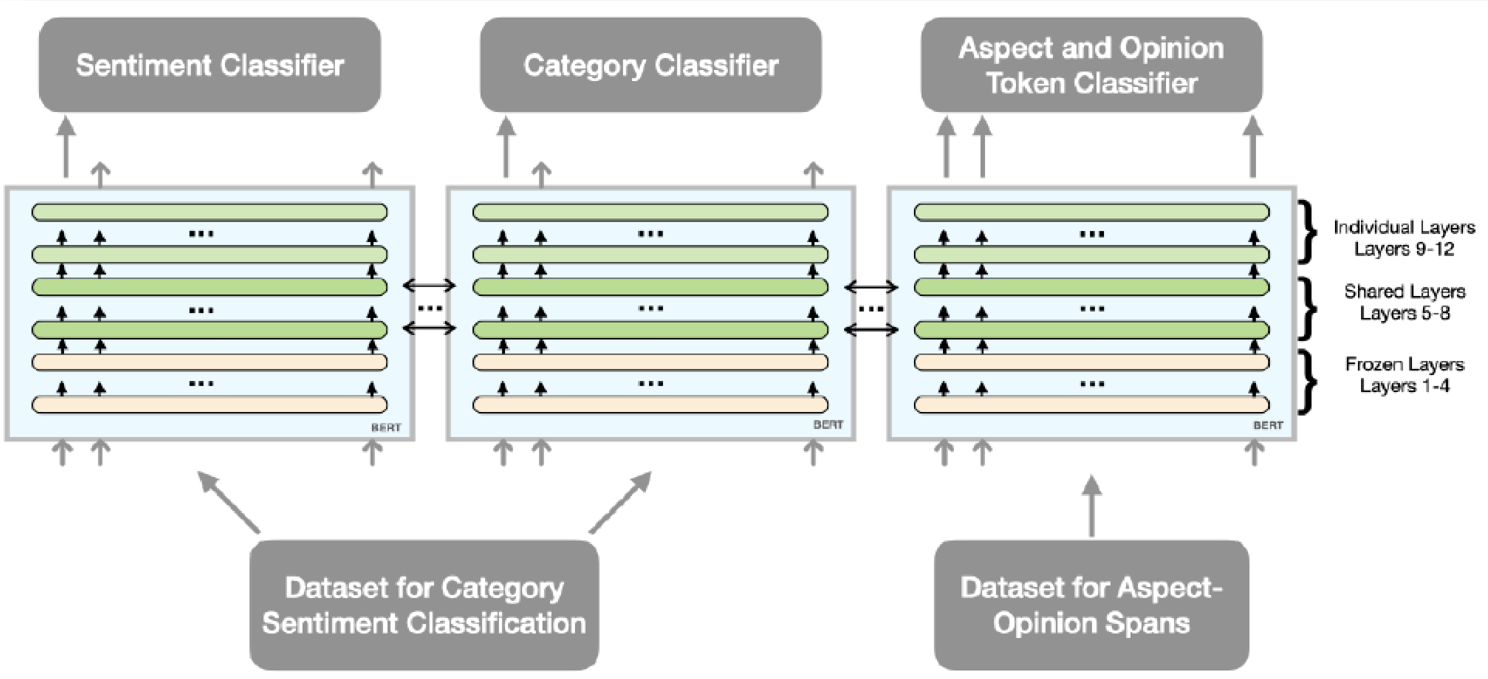
Masaki Takeo (Master student 2023-now)

Learning Simultaneously

Sentiment Analysis with Language Models

Multitask learning

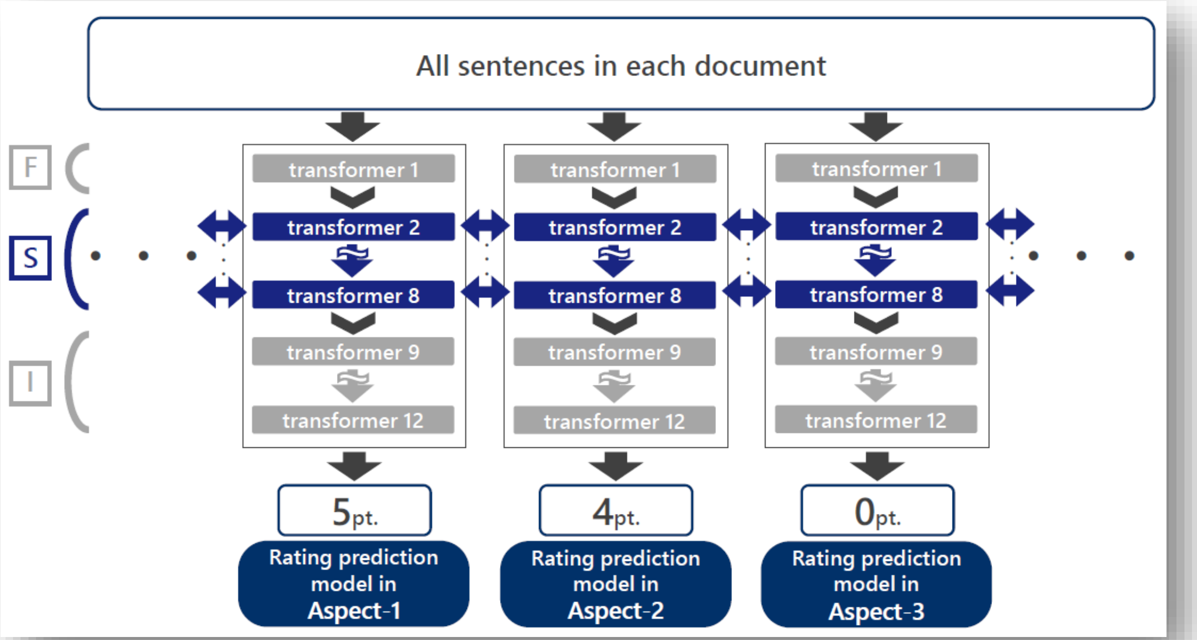
- Knowledge from similar tasks is useful
  - Positive-negative classification and opinion target detection
- Multitask learning: Sharing knowledge ≒ sharing parameters
  - In addition, it helps from a data insufficiency problem



Model F-S-I	Sent.	Cate.	Token
Each	0.54	0.17	0.51
4-4-4	0.54	0.28	0.76
1-7-4	0.55	0.30	0.76

Simultaneous learning

- Same task but different target
  - Rating prediction (☆/☆☆/☆☆☆/☆☆☆☆/☆☆☆☆☆) for seven aspects



RMSE (Lower is better)

Aspect	SVR	BERT	Ours
Addiction	1.039	0.840	0.693
Comfort	0.972	0.784	0.708
Difficulty	0.953	0.797	0.707
Graphics	0.816	0.707	0.649
Music	0.789	0.720	0.701
Originality	0.848	0.726	0.821
Satisfaction	1.058	0.830	0.715

Niraj Pahari (PhD candidate 2022-2024)

Zhu Chengcheng (Master student 2023-now)

Multilingual and Simultaneous

Sentiment Analysis with Language Models

Multilingual transformer

- If the dataset for pre-training is multilingual, the model also becomes multilingual
  - mBERT (104 languages), XLM-R (100 languages), MuRIL (16 Indian languages)
  - Current GPT is also a multilingual model
- Despite receiving no explicit information to differentiate among the languages, representation are generalized across languages
- Cross-lingual transfer
  - e.g., Fine-tuned by German, improve the accuracy for French
  - Effective for low-resource languages
    - Knowledge transfer from rich-resource languages

Note that a fine-tuned model with the target language is better if the training data are enough

Code mixing: Linguistic Behavior

- Language spoken by multilingual individual is closely related to emotion [1]
- Emotion is a driving factor for code switching behavior [2]
- Multilingual speakers have a certain language of preference for expressing their emotions [3, 4]

[1] Rajagopalan, Kanavillil. "Emotion and language politics: The Brazilian case." Journal of multilingual and multicultural development 25.2-3 (2004): 105-123.

[2] Ndubuisi-Obi, Innocent, Sayan Ghosh, and David Jurgens. "Wetin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior in Nigerian online discussions." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[3]Dewaele, Jean-Marc. Emotions in multiple languages. Basingstoke: Palgrave Macmillan, 2010.

[4]Rudra, Koustav, et al. "Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter?." Proceedings of the 2016 conference on empirical methods in natural language processing. 2016.

Research Question and Hypothesis

- Research Question
  - Do Nepali-English speakers have a preference for using native language while expressing negative sentiment in social media?
- Hypothesis I:
  - There is an association between sentiment and language proportions.
- Hypothesis II:
  - The proportion of Nepali language use is higher for negative sentences than positive sentences.

Language Identification Model

- Each token in a sentence should be given appropriate tag.
- Each token should be classified as the particular language label (lang1, lang2), named entity label (ne), ambiguous (amb) or others label (O).

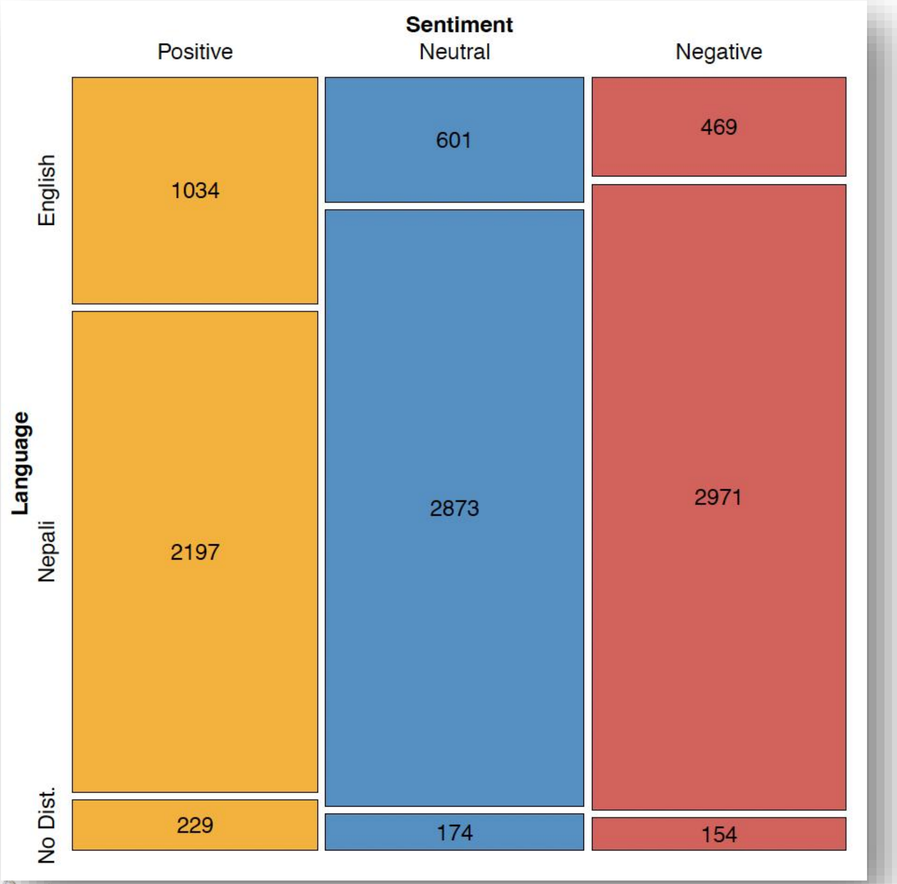
Imakara	Niraj	ga	kenkyu	naiyou	no	presentation	wo	shimasu	.
L1	NE	L1	L1	L1	L1	L2	L1	L1	O

Well-known as Sequence labeling in NLP

- Target: sentences with high Code Mixing Index (CMI)

$$CMI = \begin{cases} 100 * [1 - \frac{max(w_i)}{n-u}], & \text{if } n > u \\ 0, & \text{if } n = u \end{cases}$$

Statistical Tests



- Hypothesis I: There is an association between sentiment and language proportions
  - Chi-squared test (χ^2) is used.
  - This test checks the independence between two categorical variables.
- Hypothesis II: The proportion of Nepali language use is higher for negative sentences than positive sentences
  - Z-test for proportions is used.
  - This test checks the difference between the proportions of two samples.

WordClouds for each calss

(a) Positive (b) Neutral (c) Negative

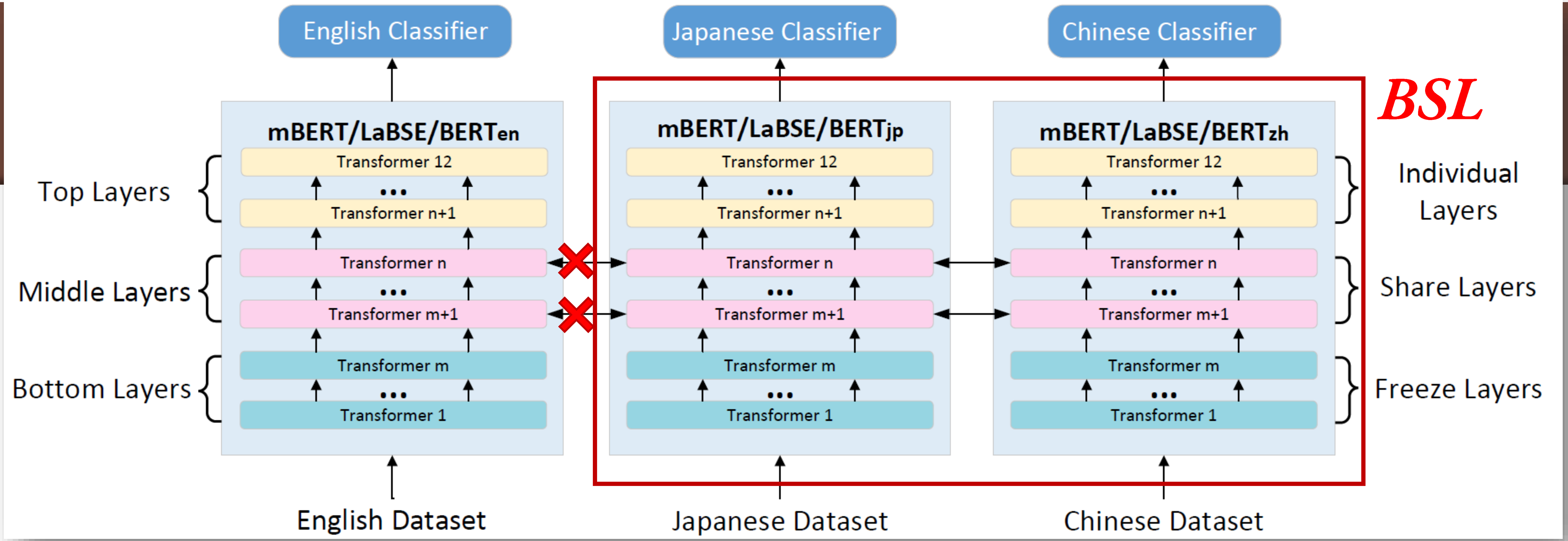
Simultaneous use of monolingual and multilingual BERTs

- A symptom prediction task on the web
  - English/Japanese/Chinese tweets (NTCIR-13, MedWeb [5])
    - Original is Japanese, and then translated into English and Chinese

<table>
<tr><th>Lang</th><th>Pseudo-tweets</th><th>Flu</th><th>Diarrhea</th><th>Hay fever</th><th>Cough</th><th>Headache</th><th>Fever</th><th>Runny nose</th><th>Cold</th></tr>
<tr><td>en</td><td>I have a fever but I don’t think it’s the kind of cold that will make it to my stomach.</td><td rowspan="3">n</td><td rowspan="3">n</td><td rowspan="3">n</td><td rowspan="3">n</td><td rowspan="3">n</td><td rowspan="3">p</td><td rowspan="3">n</td><td rowspan="3">p</td></tr>
<tr><td>ja</td><td>熱は出てるけどお腹に来る風邪じゃなさそう.</td></tr>
<tr><td>zh</td><td>虽然发烧，但是好像不是肚子着凉的感冒。</td></tr>
</table>

- Even monolingual BERTs, we can learn the models simultaneously, because the structure of each BERT is the same

[5] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. Overview of the ntcir-13: Medweb task. In NTCIR, 2017.



Model	en	ja	zh	Δ	Δ_{mono}
Baseline (STL)					
$Single_{mBERT}$	0.794	0.855	0.852		
$Single_{LaBSE}$	0.805	0.861	0.844		
$Single_{mono}$	0.838	0.856	0.873		
MSL(1-7-4)					
MSL_{mBERT}	0.834	0.850	0.856	+0.014	-0.009
MSL_{LaBSE}	**0.848**	**0.869**	0.866	+0.024	+0.005
MSL_{mono}	0.817	0.850	**0.877**	-0.008	-0.008

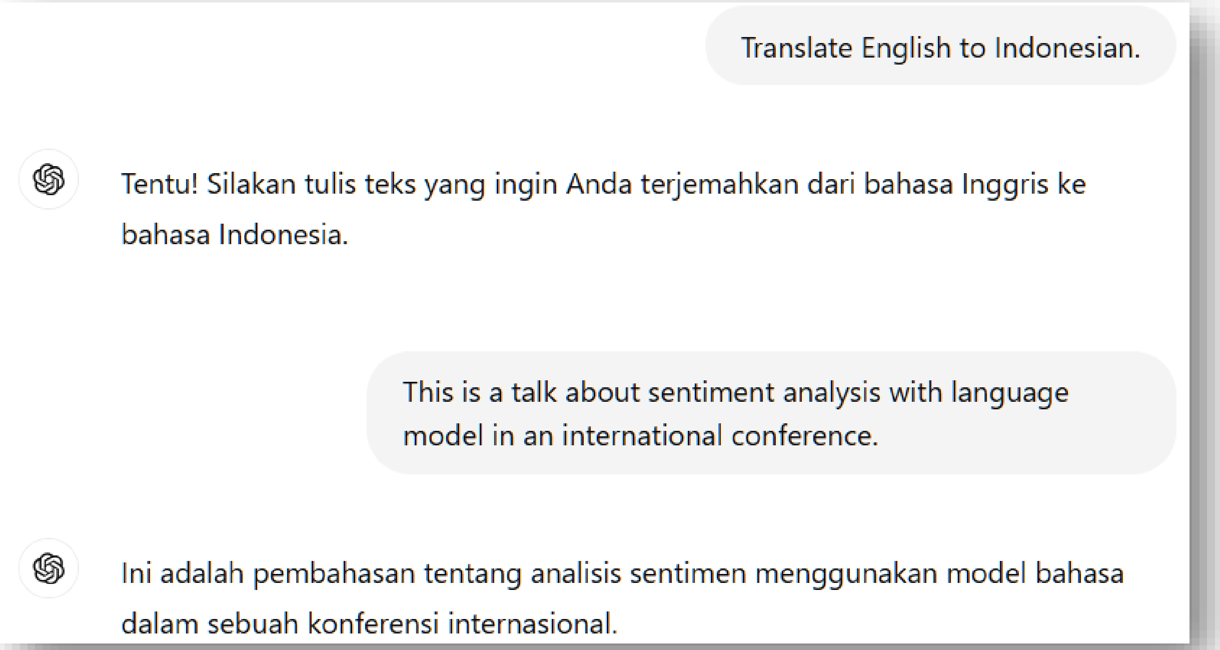
Model	ja	zh	Δ	Δ_{mono}
Baseline				
$Single_{mono}$	0.856	0.873		
MSL(1-7-4)				
MSL_{LaBSE}	0.869	0.866		+0.002
MSL_{mono}	0.850	**0.877**		-0.002
BSL(MSL without en)				
BSL_{LaBSE}	**0.873**	0.866	+0.005	**+0.005**
BSL_{mono}	0.869	0.870	**+0.006**	**+0.005**

Current large language models

Sentiment Analysis with Language Models

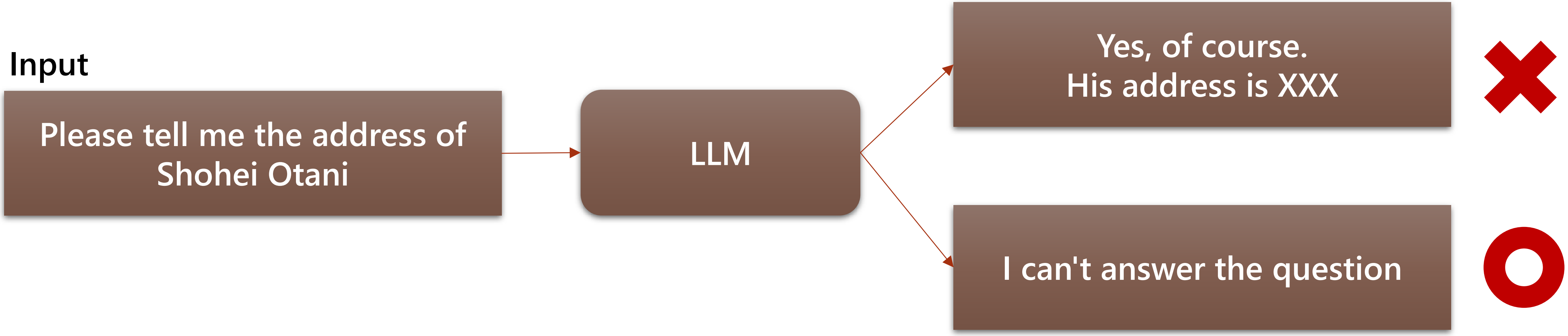
GPT：Generative Pre-trained Transformer

- Transformer-based Language model from OpenAI
  - GPT: Transformer-based Language model (Pre-trained by predicting the next word from the previous ones)
- Zero-shot
  - just the task description
- Few-shot
  - with some examples
- Basically, no parameter-tuning
  - Removing the need for task-specific architectures



Preference tuning

- Tune a model with a preference dataset
  - Preference: the fact that you like something or someone more than another thing or person from Cambridge dictionary
  - Learning whether the response about a prompt is suitable or not



RLHF

- Reinforcement Learning from Human Feedback
  - Tuning the model for a purpose with Human feedback
    - Not only ChatGPT but also other tasks; ChatGPT is just one case study as a dialogue system
  - Basic idea

1. Human judges/evaluates the output of a model
2. A reward model learns the reword from the result
3. Original language model is re-learnt by the reword model

Niraj Pahari (PhD candidate 2022-2024)

Koki Imazato (Master student 2024-now)

Approaches with LLMs

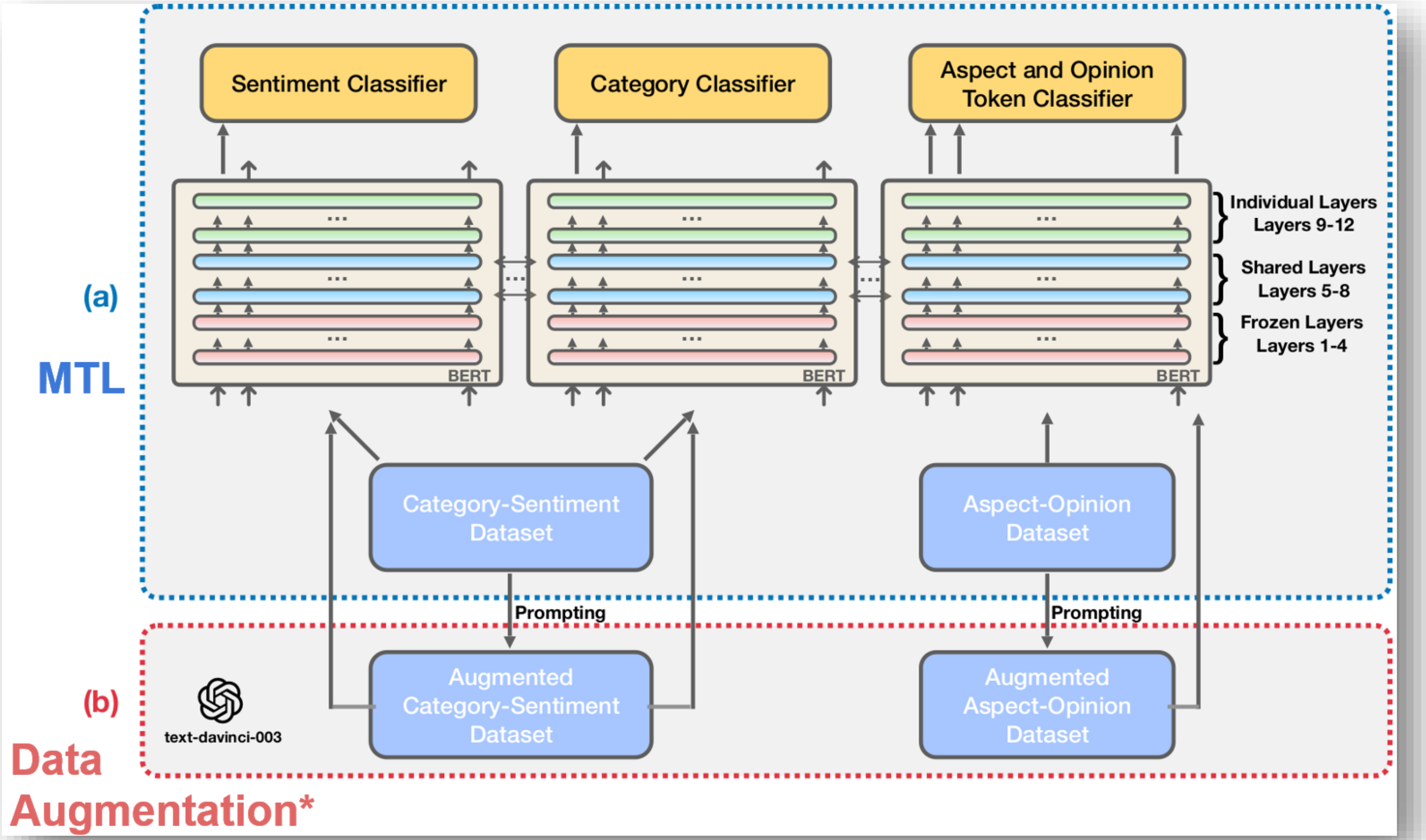
Sentiment Analysis with Language Models

Utilization of LLMs

Cat is cat, anytime

- Data augmentation
  - Effective especially image processing
    - rotating images, shifting images, and so on
  - Before LLMs, word replacing, deleting, and so on
    - Too naïve and specific issue for language: semantic drift
      - I like cats -> I dislike cats : is this Okay as the training data for sentiment analysis?
      - I like cats -> I like dogs
    - Making new data by human workers: expensive!!
- LLMs can generate fluent sentences without additional learning
  - Prompt: `Please make sentences with positive sentiment`
    - Easy-to-create with less semantic drift
    - Low-cost (a few dollars)

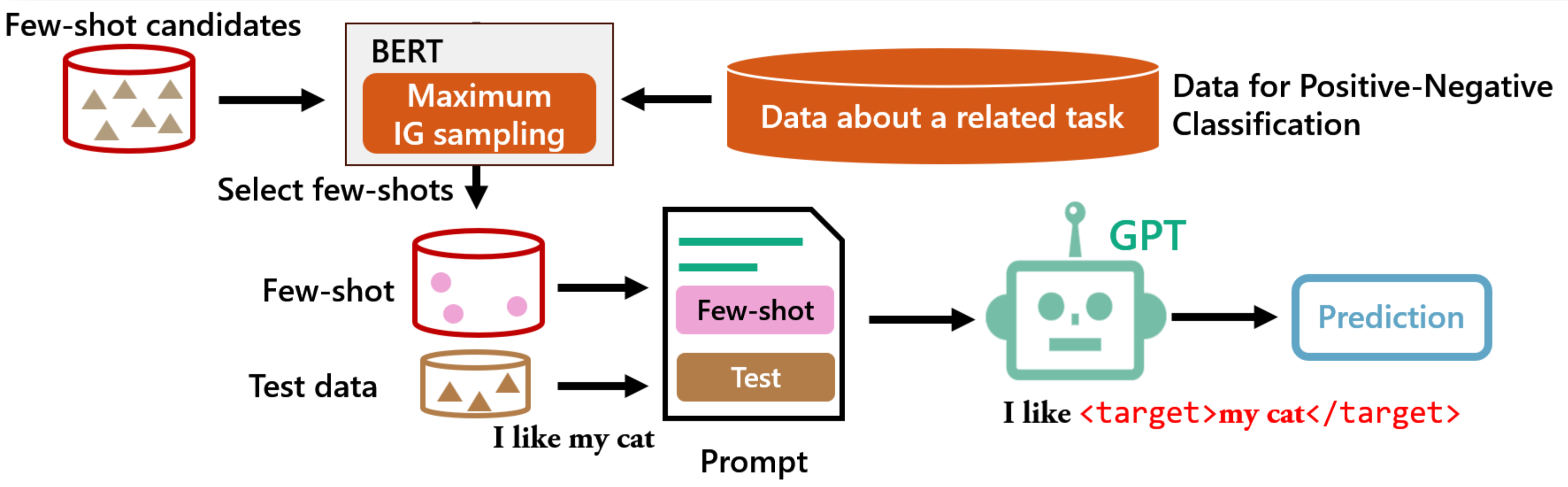
MTL with GPT-based data augmentation



Model	Sent.	Cate.	Token	Ave.
Single	0.54	0.17	0.51	0.41
Single+ Aug	0.59	0.35	0.75	0.56
MTL+ Aug	0.55	0.36	0.72	0.54
MTL+ Cate.&Sent. Aug.	0.60	0.37	0.74	0.57

Few-shot selection

- Better few-shot, better performance for the target task
- Few-shot selection model based on related tasks



$$IG(Y|x_i) = H(Y) - H(Y|x_i)$$

The assumption is that BERT understands sentiment analysis easier when the evaluation target is explicit.

Selection	F1
Random	0.481
Ours	0.513

Wang Quan (Master student 2014-2015)

Are LLMs the best tool?

Sentiment Analysis with Language Models

LLMs as a tool

- LLMs are incredibly powerful tools
  - We can’t cut wood with a box cutter, but we can do it with a chainsaw.
- However, do we need a chainsaw for cutting a paper?

Everything comes in handy when used right

Just counting!

Review analysis on culture difference

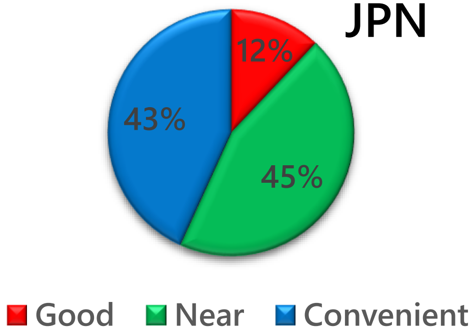
Translate into English, then analyze them

- Hotel review analysis of Japanese and Chinese

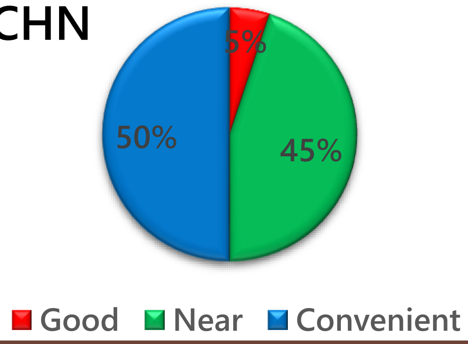
Word	Freq JPN	Expression and Freq	Freq CHN	Expression and Freq
station	3864	good(60), great(25), excellent(3), close(179), nearest(147), helpful(14), useful(32), convenient(269)	1953	good(5), excellent(2), close(52), nearest(10), convenient(69)

- Japanese guests tend to care about the location to the station
- However, the ratios for expressions with "station" remain consistent

JPN



CHN



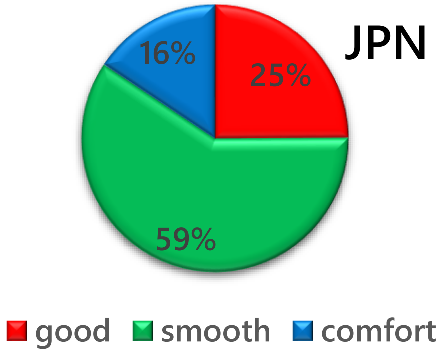
Review analysis on culture difference (cont.)

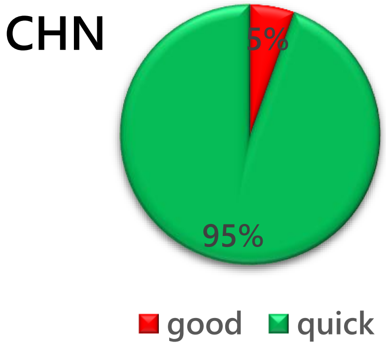
Translate into English, then analyze them

- Hotel review analysis of Japanese and Chinese

Word	Freq JPN	Expression and Freq	Freq CHN	Expression and Freq
check-in	1221	good(64),great(6), smooth(88), quickly(24), early(16), efficient(39), helpful(4), comfortable(40)	1953	great(5), quickly(31), early(17), simple(9), fast(12), smooth(7), efficient(10)

- Japanese guests evaluate the politeness of the staff during check-in, while Chinese guests tend to prioritize the speed of the check-in process.





Review analysis on culture difference (cont.)

Translate into English, then analyze them

- Hotel review analysis of Japanese and Chinese

Word	Freq JPN	Expression and Freq	Freq CHN	Expression and Freq
coffee	758	excellent(33), good(13), free(47), provided(5), available(3), delicious(46), plenty(24)	255	good(6), great(3), wonderful(1), supply(1), free(16), available(8), delicious(3), aromatic(2)

Coffee for JAPANESE!!

Note that expressions related to quantity, such as *plenty*, appear only in Japanese

Thanks for my lab members

Summary

Sentiment Analysis with Language Models

Conclusions

- What is a language model?
    - Traditional and current language models
- Sentiment analysis with language models
    - BERT with some additional knowledge/resources
    - Learning simultaneously
        - Multitask learning
        - Multilingual situation
    - Use of Large Language Models
        - Data augmentation
        - Few-shot selection
    - LLMs are just a tool for analysis

Papers are on my web page
This slides will appear on the web page

List of paper in this presentation

- Some explanations in this presentation are not published yet, but the main content appears on
  - Satoshi Hiai and Kazutaka Shimada Sarcasm Detection Using RNN with Relation Vector . he International Journal of Data Warehousing and Mining (IJDWM), Volume 15, Issue 4, pp. 66-78, 2019.
  - Zhenming Li and Kazutaka Shimada. Combination and Knowledge Extension of Pre-trained Language Model for Offensive Language Detection. 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 82-87, 2023.
  - Masaki Takeo and Kazutaka Shimada. Rating Prediction of Multi-aspect Reviews Using Simultaneous Learning. 2023 International Conference on Asian Language Processing (IALP), pp. 358-363, 2023.
  - Niraj Pahari and Kazutaka Shimada. Layer Configurations of BERT for Multitask Learning and Data Augmentation. Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 28. No. 1, pp. 29-40, 2024.
  - Niraj Pahari and Kazutaka Shimada. Language Preference for Expression of Sentiment for Nepali-English Bilingual Speakers on Social Media. Proceedings of Workshop CALCS (EMNLP2023), pp. 23-32, 2023.
  - Zhu Chencheng, Niraj Pahari, and Kazutaka Shimada. Multilingual Symptom Prediction by Simultaneous Learning using BERT. 2023 International Conference on Asian Language Processing (IALP), pp. 100-105, 2023.
  - Koki Imazato and Kazutaka Shimada. Automatic Few-shot Selection on In-Context Learning for Aspect Term Extraction. 2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 15-20, 2024.
- This research is published by Japanese version only
  - Quan Wang and Kazutaka Shimada. Detection of differences on evaluation points between Japanese and Chinese. IEICE Technical Report NLC2014-49, 2015.