A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification

Sosuke Nishikawa[1,2], Ikuya Yamada[2,4], Yoshimasa Tsuruoka[1] and Isao Echizen[1,3]

[1]The University of Tokyo, Japan
[2]Studio Ousia, Japan
[3]National Institute of Informatics, Japan
[4]RIKEN, Japan

Outline

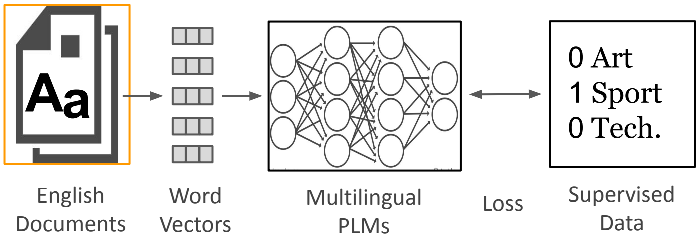
- Background & Motivation
- Proposed Method
- Experiments
- Analysis
- Conclusion

Outline

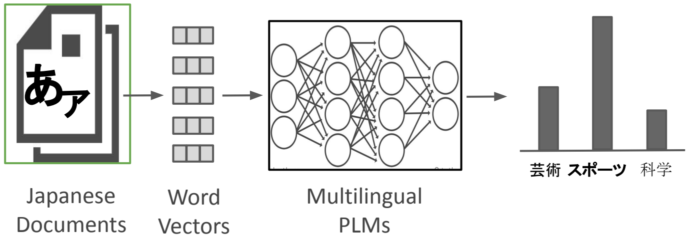
- **Background & Motivation**
- Proposed Method
- Experiments
- Analysis
- Conclusion

Zero-Shot Cross-lingual Text Classification

Training: annotated resource-rich language

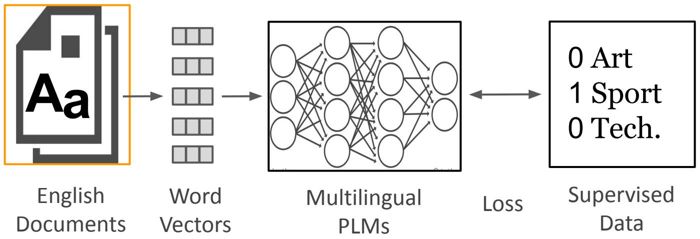


Inference: target language

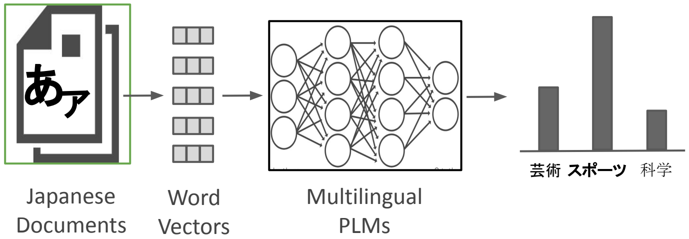


Zero-Shot Cross-lingual Text Classification

Training: annotated resource-rich language



Inference: target language



→ Substantial progress in cross-lingual transfer learning has been made using multilingual pre-trained language models (PLMs) such as M-BERT

Limitations of existing methods 1

Multilingual PLMs do not always perform well in cross-lingual transfer in the following cases [Conneau et al., 2020, Lauscher et al., 2020]:

- ➢ Transfer to a typologically different language (e.g., SOV languages -> SVO languages)
- ➢ Transfer to a language with a small amount of pre-training data

Limitations of existing methods 2

There are many methods to further train multilingual PLMs using unlabeled text data in certain target languages [Eisenschlos et al., 2019, Conneau et al., 2019, Lai et al., 2019]

- ➢ These methods require extra training using additional text for each language
- ➢ These methods work well only on a single target language

Limitations of existing methods 2

There are many methods to further train multilingual PLMs using unlabeled text data in certain target languages [Eisenschlos et al., 2019, Conneau et al., 2019, Lai et al., 2019]

- ➢ These methods require extra training using additional text for each language
- ➢ These methods work well only on a single target language

→ Our work alleviates these limitations

by using **knowledge base entities as input features**

Entity as training resource

Knowledge base (KB) entities as input features has following advantages:

- Knowledge Base (KB) entities can capture unambiguous semantics in documents and have been used to address text classification [Song et al., 2016, Yamada and Shindo, 2019]
- KB entities are defined independently of languages [Calixto et al., 2021]

Outline

- ➢ Background & Motivation
- ➢ **Proposed Method**
- ➢ Experiments
- ➢ Analysis
- ➢ Conclusion

Key Idea

Extract language-agnostic Wikidata entities

and use them as features for cross-lingual text classification

引け前の台湾株式市場で、加権指数が3.28%急落した。フローカーらによると、工業株に売りが集中したため、という。大引け前10分現在加権指数は278.07ポイント急落し、8207.59。

Japanese document about
the Taiwanese stock market

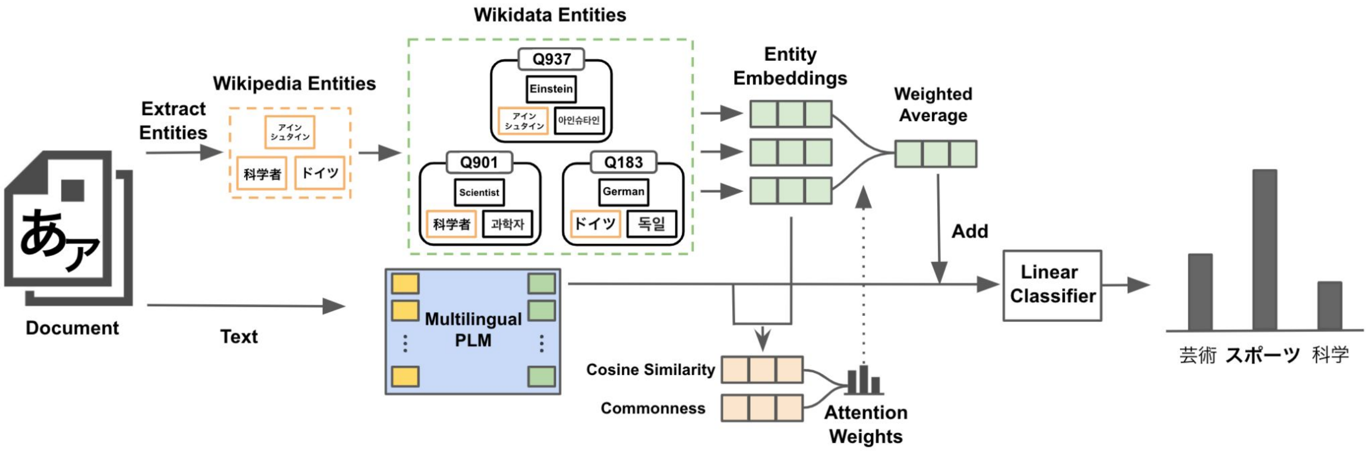
Stock certificate
（Q855349）

Share price
（Q1020013）

Taiwan Capitalization Weighted Stock Index
（Q448773）

Language-agnostic Wikidata entities

Multilingual Bag-of-Entities Model (M-BoE)

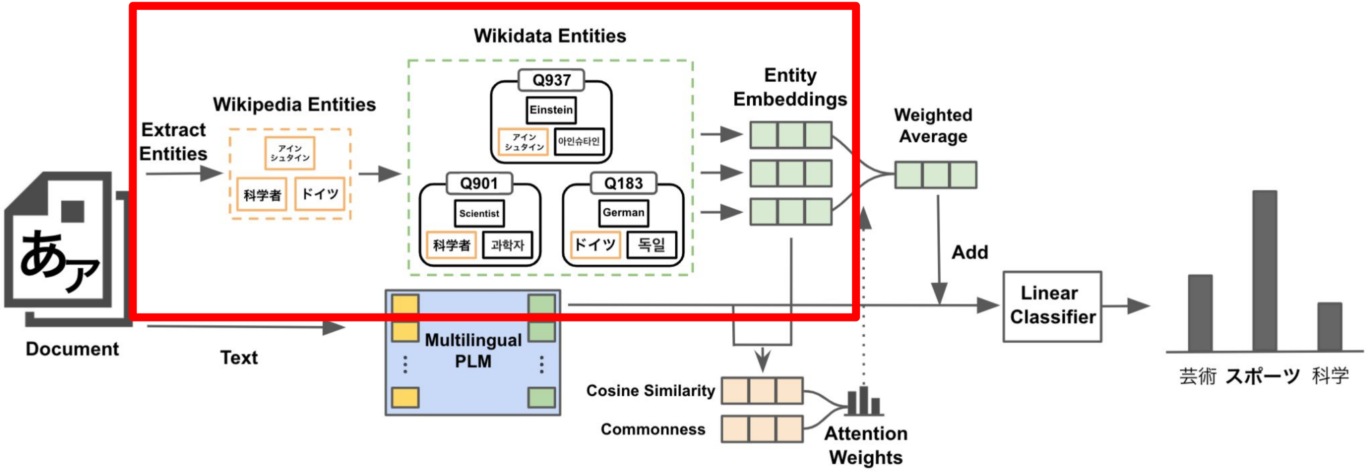


M-BoE is a simple extension of multilingual PLMs

by **using Wikidata entities as additional input features**

Multilingual Bag-of-Entities Model (M-BoE)

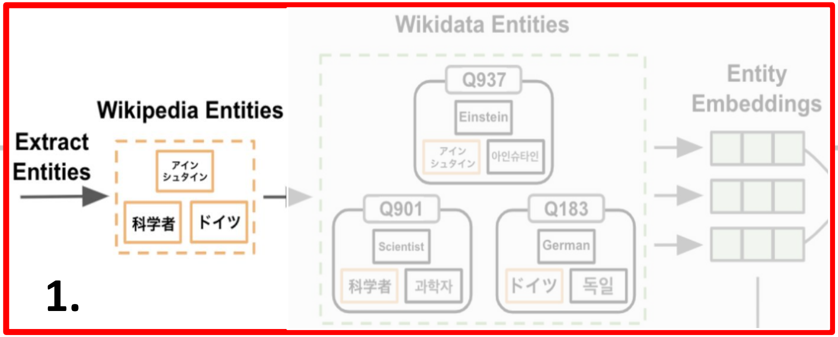
Entity Preprocessing



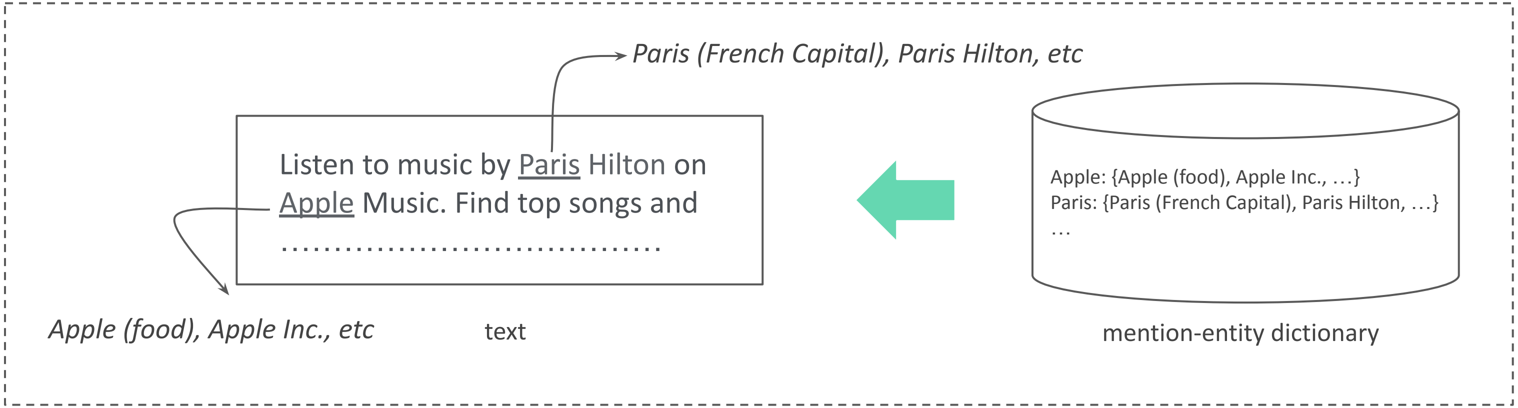
M-BoE is a simple extension of multilingual PLMs

by **using Wikidata entities as additional input features**

Entity Preprocessing

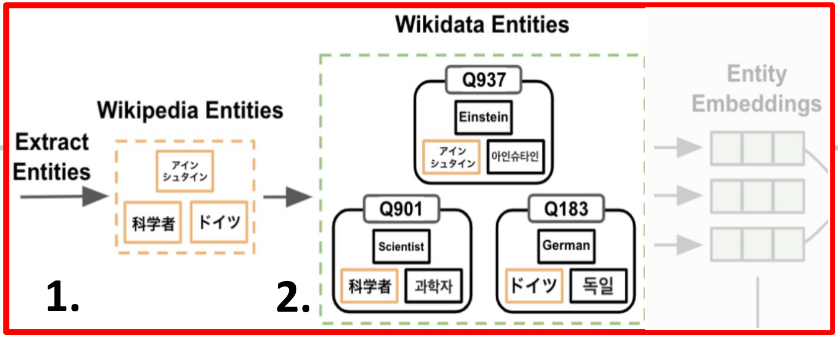


1. Detect all possible referent Wikipedia entities for each detected entity name using the mention-entity dictionary

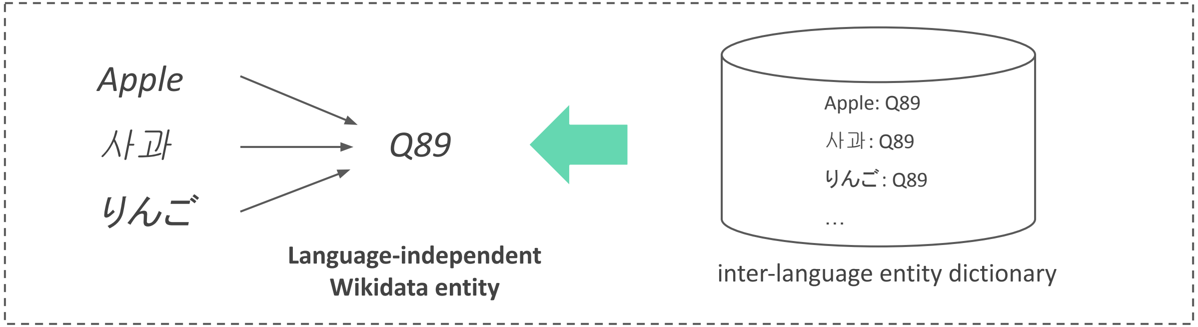


Illustration

Entity Preprocessing

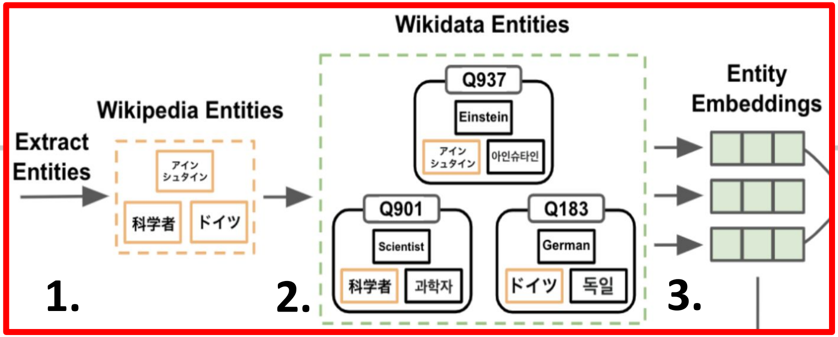


1. Detect all possible referent Wikipedia entities for each detected entity name using the mention-entity dictionary
2. Convert detected Wikipedia entities to Wikidata entities using the inter-language entity dictionary



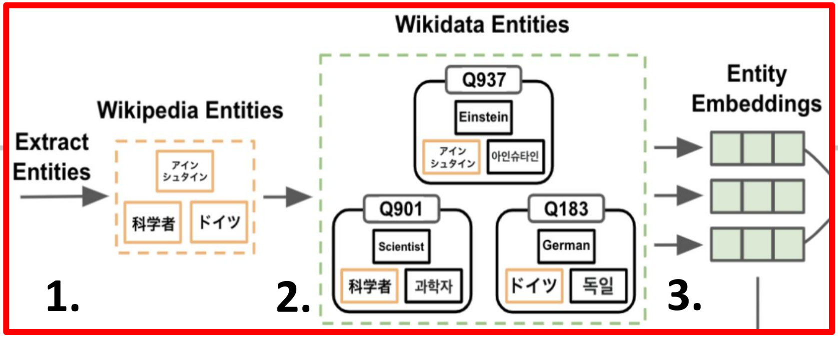
Illustration

Entity Preprocessing



1. Detect all possible referent Wikipedia entities for each detected entity name using the mention-entity dictionary
2. Convert detected Wikipedia entities to Wikidata entities using the inter-language entity dictionary
3. Assign an entity embedding for each Wikidata entity

Entity Preprocessing

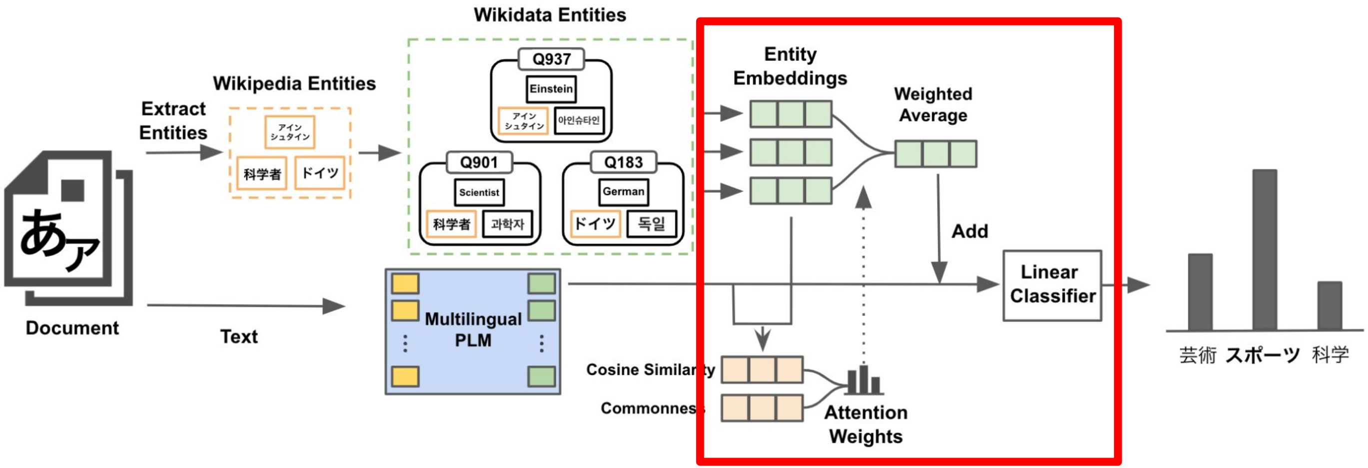


1. Detect all possible referent Wikipedia entities for each detected entity name using the mention-entity dictionary e.g., "Apple" -> "Apple Inc."
2. Convert detected Wikipedia entities to Wikidata entities using the inter-language entity dictionary e.g., "Apple Inc." -> "Q312" (language-independent)
3. Assign an entity embedding for each Wikidata entity

→ This enables **entities in multiple languages to be represented using shared embeddings**

Multilingual Bag-of-Entities Model (M-BoE)

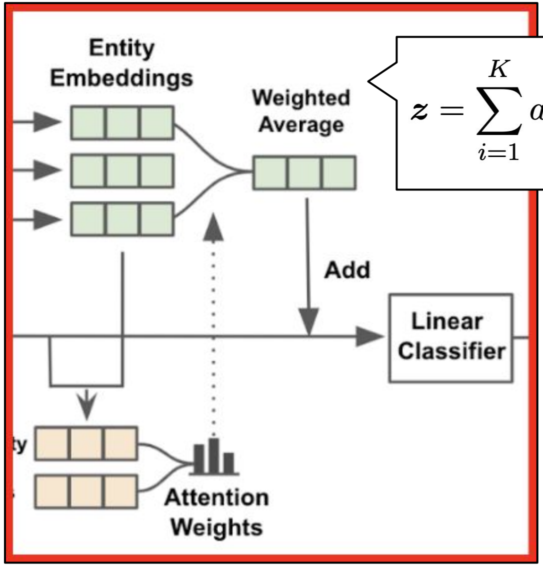
Entity-based representation



M-BoE is a simple extension of multilingual PLMs

by **using Wikidata entities as additional input features**

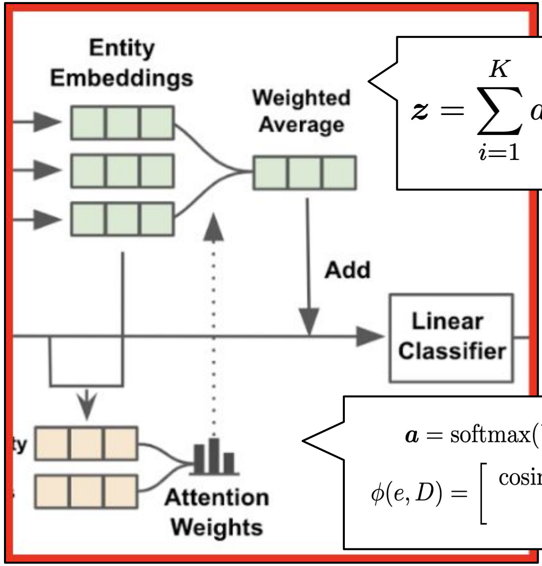
Entity-based Text Representation



$$z = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v}_{e_i}$$

- The entity-based text representation **z** is computed as the weighted average of entity embeddings **v**

Entity-based Text Representation



$$z = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v}_{e_i}$$

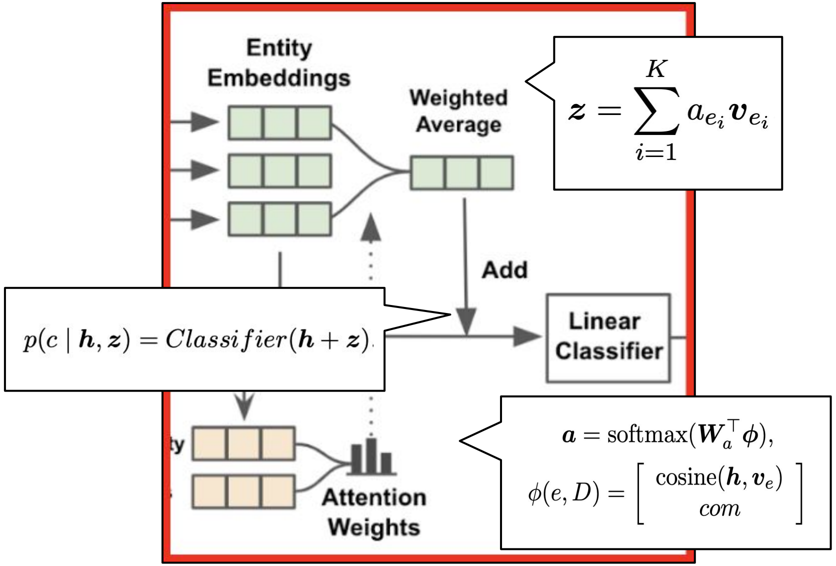
$$\boldsymbol{a} = \mathrm{softmax}(\boldsymbol{W}_a^\top \boldsymbol{\phi}),$$

$$\phi(e, D) = \begin{bmatrix} \mathrm{cosine}(\boldsymbol{h}, \boldsymbol{v}_e) \\ com \end{bmatrix}$$

[Yamada et al., 2019, Peters et al., 2019]

- ➢ The entity-based text representation **z** is computed as the weighted average of entity embeddings **v**
- ➢ The weights are computed by the attention mechanism
  - **h**: text-based representation
  - Com: Probability that an entity name refers to an entity in KB

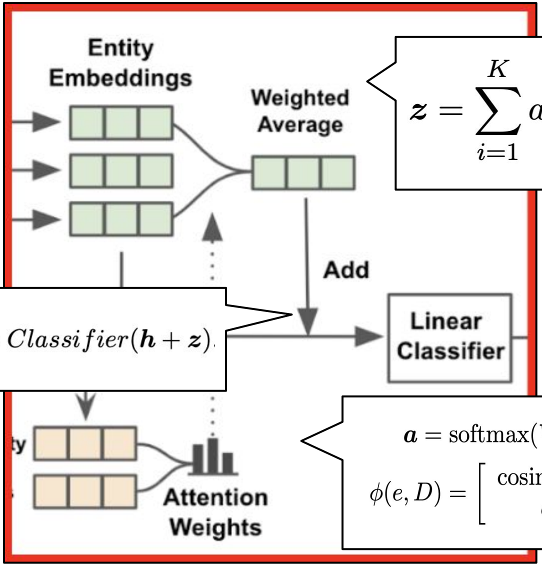
Entity-based Text Representation



[Yamada et al., 2019, Peters et al., 2019]

- The entity-based text representation **z** is computed as the weighted average of entity embeddings **v**
- The weights are computed by the attention mechanism
  - **h**: text-based representation
  - Com: Probability that an entity name refers to an entity in KB
- The sum of these embeddings is fed into a linear classifier

Entity-based Text Representation



$$z = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v}_{e_i}$$

$$p(c \mid \boldsymbol{h}, \boldsymbol{z}) = Classifier(\boldsymbol{h} + \boldsymbol{z})$$

$$\boldsymbol{a} = \mathrm{softmax}(\boldsymbol{W}_a^\top \boldsymbol{\phi}),$$

$$\phi(e, D) = \begin{bmatrix} \mathrm{cosine}(\boldsymbol{h}, \boldsymbol{v}_e) \\ com \end{bmatrix}$$

[Yamada et al., 2019, Peters et al., 2019]

- The entity-based text representation **z** is computed as the weighted average of entity embeddings **v**
- The weights are computed by the attention mechanism
  - **h**: text-based representation
  - Com: Probability that an entity name refers to an entity in KB
- The sum of these embeddings is fed into a linear classifier

→ M-BoE can automatically select **the entities that are effective in solving the classification task**

Model features of M-BoE

- ✓ M-BoE is a simple extension of a PLM and does not modify its internal architecture
- ✓ M-BoE boosts performance in multiple languages simultaneously by training only a single model
- ✓ M-BoE does not need expensive pre-training and additional text data in the target languages

Outline

- ➢ Background & Motivation
- ➢ Proposed Method
- ➢ **Experiments**
- ➢ Analysis
- ➢ Conclusion

Experimental setting

Base models:	multilingual BERT, XLM-R
Classification Datasets:	MLDoc, TED-CLDC, and SHINRA2020-ML
Compared models:	LASER [Artetxe et al., 2019] , MultiCCA [Schwenk et al., 2018])
Entity embeddings:	Wikipedia2Vec [Yamada et al., 2020] embeddings trained from English Wikipedia dump
Evaluation setting:	Train in English and evaluate in other languages

Results

Model	en	fr	de	ja	zh	it	ru	es	target avg.
MultiCCA (Schwenk and Li, 2018)	92.2	72.4	81.2	67.6	74.7	69.4	60.8	72.5	71.2
LASER (Artetxe and Schwenk, 2019)	89.9	78.0	84.8	60.3	71.9	69.4	67.8	77.3	72.8
M-BERT	94.0	79.4	75.1	69.3	68.0	67.1	65.3	75.2	71.4 ± 1.4
+M-BoE	94.1	84.0	76.9	71.1	72.2	70.0	68.9	75.5	74.1 ± 0.7
XLM-R	94.4	84.9	86.7	78.5	85.2	73.4	71.3	81.5	80.2 ± 0.5
+M-BoE	**94.6**	**86.4**	**88.9**	**80.0**	**87.4**	**75.6**	**73.7**	**83.2**	$\mathbf{82.2 \pm 0.6}$

Table 2: Classification accuracy for topic classification on MLDoc dataset; "target avg." indicates average scores for target languages.

Model	en	fr	de	it	ru	es	ar	tr	nl	pt	pl	ro	target avg.
M-BERT	51.6	47.7	43.9	50.6	47.9	53.1	41.3	44.2	49.4	46.2	45.1	45.4	47.1 ± 1.4
+M-BoE	**52.9**	49.5	46.2	**53.3**	49.2	**54.7**	44.7	49.1	51.0	47.6	47.7	48.2	49.6 ± 1.1
XLM-R	51.5	49.5	49.7	48.7	48.3	51.2	45.6	51.3	48.8	46.3	48.3	48.4	49.1 ± 1.8
+M-BoE	51.7	**50.0**	**53.8**	51.3	**52.3**	52.9	**50.5**	**53.1**	**52.0**	**49.3**	**50.5**	**49.6**	$\mathbf{51.8 \pm 0.9}$

Table 3: F1 score for topic classification on TED-CLDC dataset.

- ✓ M-BoE outperformed state-of-the-art methods for a diverse range of languages
- ✓ Observed similar trends in SHINRA2020-ML dataset

Outline

- ➢ Background & Motivation
- ➢ Proposed Method
- ➢ Experiments
- ➢ **Analysis**
- ➢ Conclusion

Impact of each component in M-BoE

Analyzed the impact on the performance of each component in the M-BoE model

Setting	M-BoE (M-BERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9
Baseline	71.4	80.2

Table 5: Results of analysis of our model on MLDoc.

Impact of each component in M-BoE

Analyzed the impact on the performance of each component in the M-BoE model

Setting	M-BoE (M-BERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9
Baseline	71.4	80.2

Table 5: Results of analysis of our model on MLDoc.

✓ Our attention mechanism and its features are effective

Impact of each component in M-BoE

Analyzed the impact on the performance of each component in the M-BoE model

Setting	M-BoE (M-BERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9
Baseline	71.4	80.2

Table 5: Results of analysis of our model on MLDoc.

✓ Our attention mechanism and its features are effective

✓ Wikipedia2Vec is the most effective compared with random and knowledge graph (KG) embeddings

Impact of each component in M-BoE

Analyzed the impact on the performance of each component in the M-BoE model

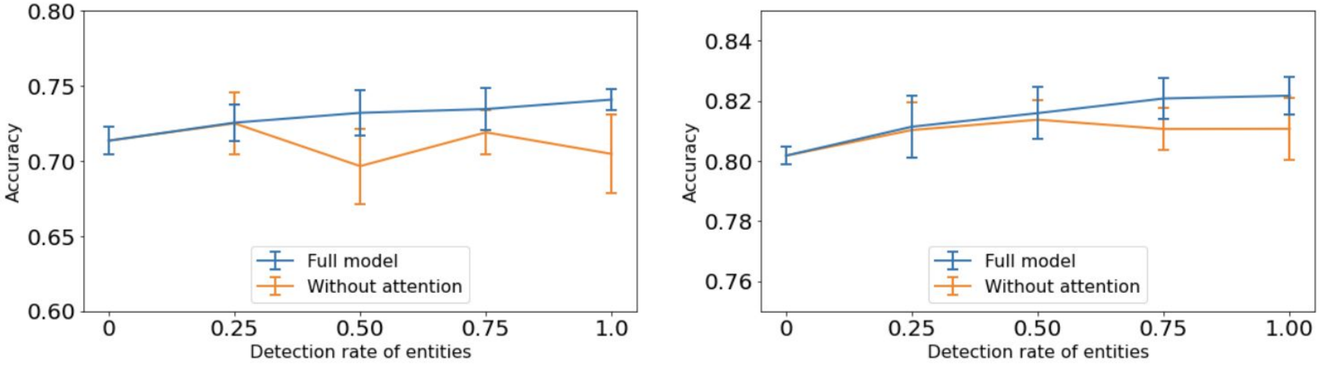
Setting	M-BoE (M-BERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9
Baseline	71.4	80.2

Table 5: Results of analysis of our model on MLDoc.

- ✓ Our attention mechanism and its features are effective
- ✓ Wikipedia2Vec is the most effective compared with random and knowledge graph (KG) embeddings
- ✓ Our entity detection method outperformed the commercial multilingual entity linking system

Impact of the number of entities in M-BoE

Examined the performance impact of the number of detected Wikidata entities

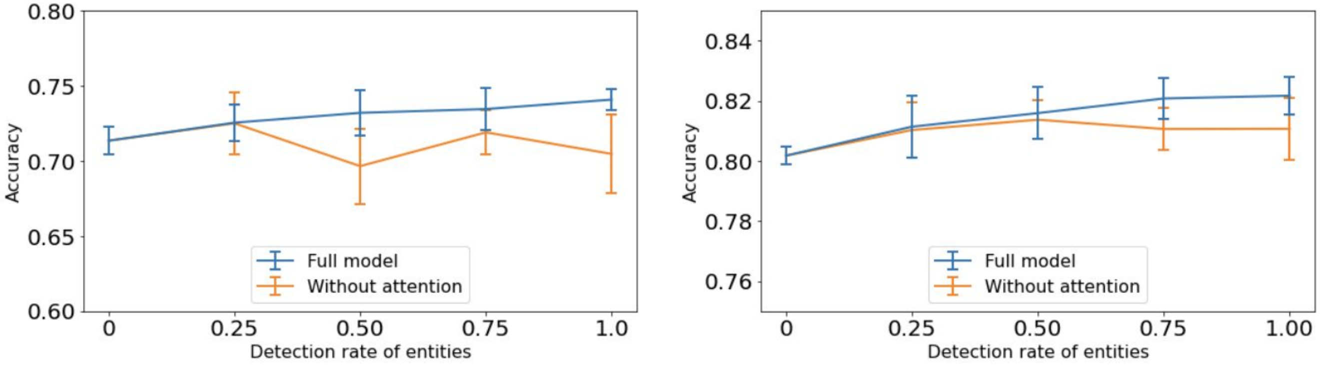


(a) M-BoE (M-BERT) (b) M-BoE (XLM-R)

Figure 2: Classification accuracy for each entity detection rate using MLDoc dataset.

Impact of the number of entities in M-BoE

Examined the performance impact of the number of detected Wikidata entities



(a) M-BoE (M-BERT) (b) M-BoE (XLM-R)

Figure 2: Classification accuracy for each entity detection rate using MLDoc dataset.

- ✓ The higher the entity detection rate, the better the performance of the full model
- ✓ Attention mechanism is important for this consistent improvement

Qualitative analysis

Examined the influential entities that were assigned the largest attention weights

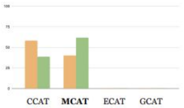
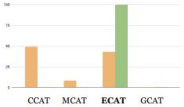
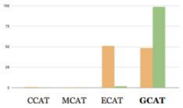
Language	Document	Label	Probability distribution M-BERT M-BoE	Top three entities
Ja	[台北 2日 ロイター] 引け前の台湾株式市場で、加権指数が3.28%急落した。フローカーらによると、工業株に売りが集中したため、という。大引け前10分(0350gmt)現在、加権指数は278. 07ポイント(3. 28%)急落し、8207. 59。売買代金は、1090億台湾ドル。	**MCAT** **(Markets)**		**"Stock certificate" "Share price" "Taiwan Capitalization Weighted Stock Index"**
Zh	〔路透社東京19日電〕日本大藏省一顧問小組週四促請大藏省取消目前只允許被授權外匯銀行進行外匯交易的管制,完全開放外匯市場交易資格的限制. 這項限制的取消將使投資人進出外匯市場更為容易;此外,銀行業也可藉此增進競爭力,並促進市場的流動性及活絡匯市的交易.(完)	**ECAT** **(Economics)**		**"Ministry of the Treasury" "Financial transaction" "Competition (economics)"**
Ru	москва, 17 мар (рейтер) - президент рф борис ельцин подписал федеральные законы о внесении изменении и дополнении в статьи 100 и 110 закона рф "о государственных пенсиях в рф", сообщила пресс-служба президента рф. статья 100 закона излагается в следующеи редакции: "в заработок для исчисления пенсии включаются все виды выплат (дохода), полученных в связи с выполнением работы, предусмотренной статьеи 89 закона, на которые начисляются страховые взносы в пенсионный фонд рф". пресс-служба президента рф сообщила, что виды выплат, на которые не начисляются страховые взносы в пенсионный фонд рф, определяются правительством рф.	**GCAT** **(Government Social)**		**"Federal law" "Pension Fund of the Russian Federation" "Kremlin Press Secretary"**

Figure 4: Example results for MLDoc. “Top three entities” indicates the three most influential entities selected by attention mechanism.

✓ Our model identified the entities that were highly relevant to the document

Conclusion

- We proposed M-BoE to improve zero-shot cross-lingual text classification by injecting language-agnostic features of Wikidata entities to multilingual PLMs
- M-BoE achieved state-of-the-art results on three cross-lingual text classification tasks
- We plan to evaluate our model on low-resource languages and a variety of natural language processing tasks

Thank you for listening