Leveraging Unlabeled Text: Data-Centric Approaches to Improve NLP Training

QIYU WU

INTERN AT CAL SECTION 4, SONY

PH.D. STUDENT AT TSURUOKA LAB

THE UNIVERSITY OF TOKYO

Immense Information in Unlabeled Text

<u>**Unlabeled text**</u> is everywhere

Large scale
- Wiki-40B: **2.9M** pages for English
- LM1B: **30M** sentences of news comments
- C4: **360M** web documents
- …

Knowledge base, e.g., Wikipedia

Non-experts, e.g., Twitter

LLMs, e.g., ChatGPT

Success with unlabeled text: Large language model, ChatGPT…

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT ↗ Read about ChatGPT Plus

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Dataset Matters for NLP Model

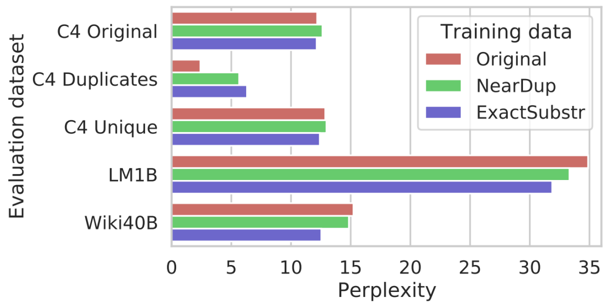
Language model can be influenced by the dataset in several aspects:

- Duplication in the dataset
- Input format
- …

We do need someone working on the data side

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Table credit to RoBERTa (Liu et al. 2019)



Deduplicating Training Data Makes Language Models Better (Lee et al. ACL 2022)

On the De-duplication of LAION-2B

Ryan Webster
Unicaen
ryan.webster@unicaen.fr

Loic Simon
ENSICAEN
loic.simon@ensicaen.fr

Julien Rabin
Unicaen
julien.rabin@unicaen.fr

Frederic Jurie
Unicaen
frederic.jurie@unicaen.fr

"Roughly 700 million, or about a third of LAION-2B's images, are duplicates" (Webster, et al. 2023)

Beyond Models: Data-Centric Approaches

Large-scale unlabeled text produce a wealth of information, but still imperfect.

My research focus lies in **better utilization of unlabeled text to improve NLP models**

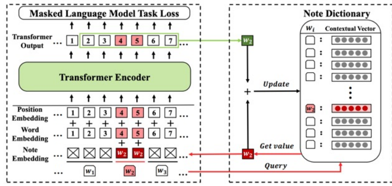
Dataset construction from
- Collection
- Input format

Dataset augmentation

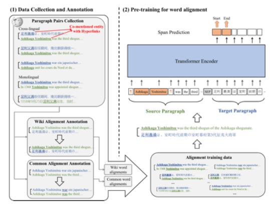
Data quality

…

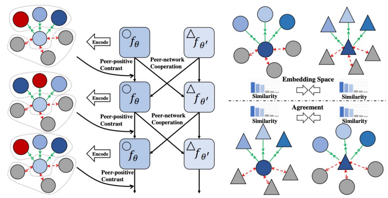
Utilize cross-sentence signal to address rare words issue in language model pre-training(Wu et al. ICLR 2021)



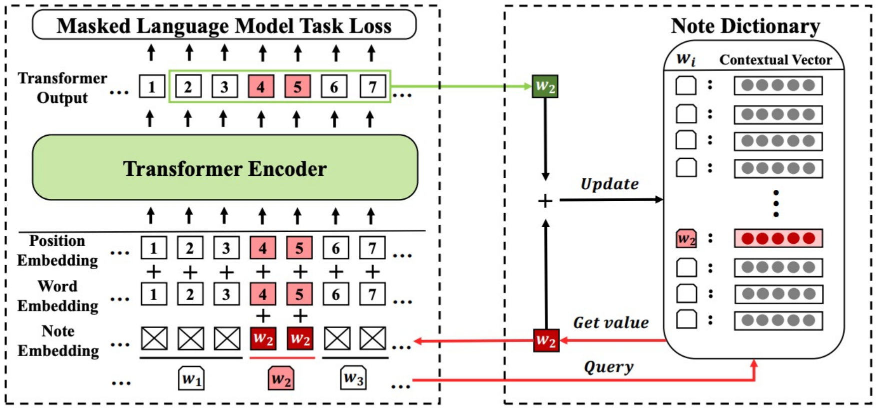
Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training (Wu et al. ACL 2023)



Unsupervised data augmentation for sentence embedding by contrastive learning(Wu et al. EMNLP 2022)



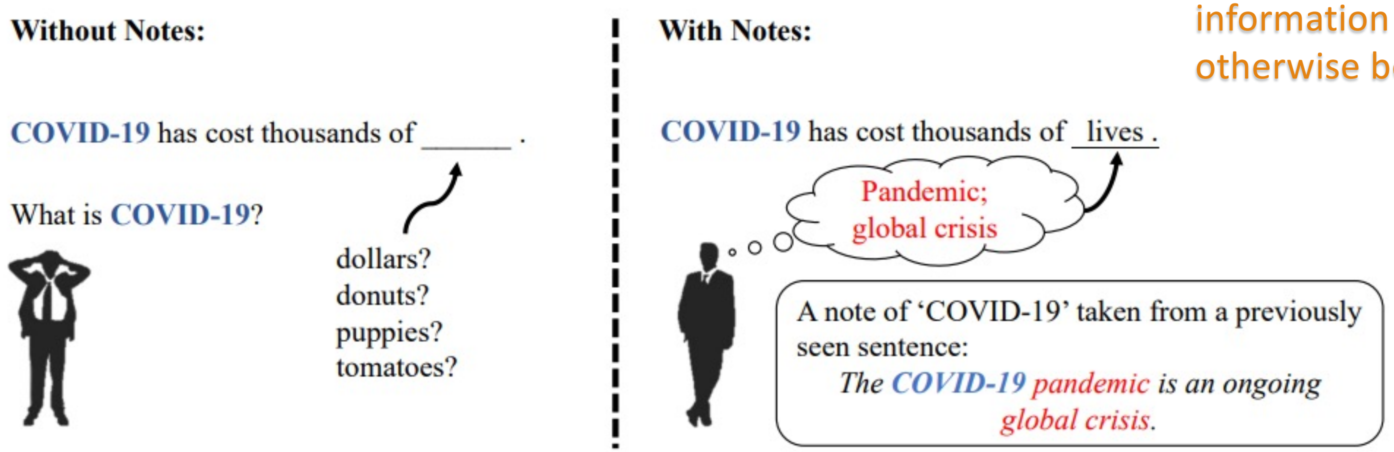
Beyond Models: Data-Centric Approaches



Utilize cross-sentence signal to address rare words issue in language model pre-training(Wu et al. ICLR 2021)

Rare words make inputs noisy, and slow down language training

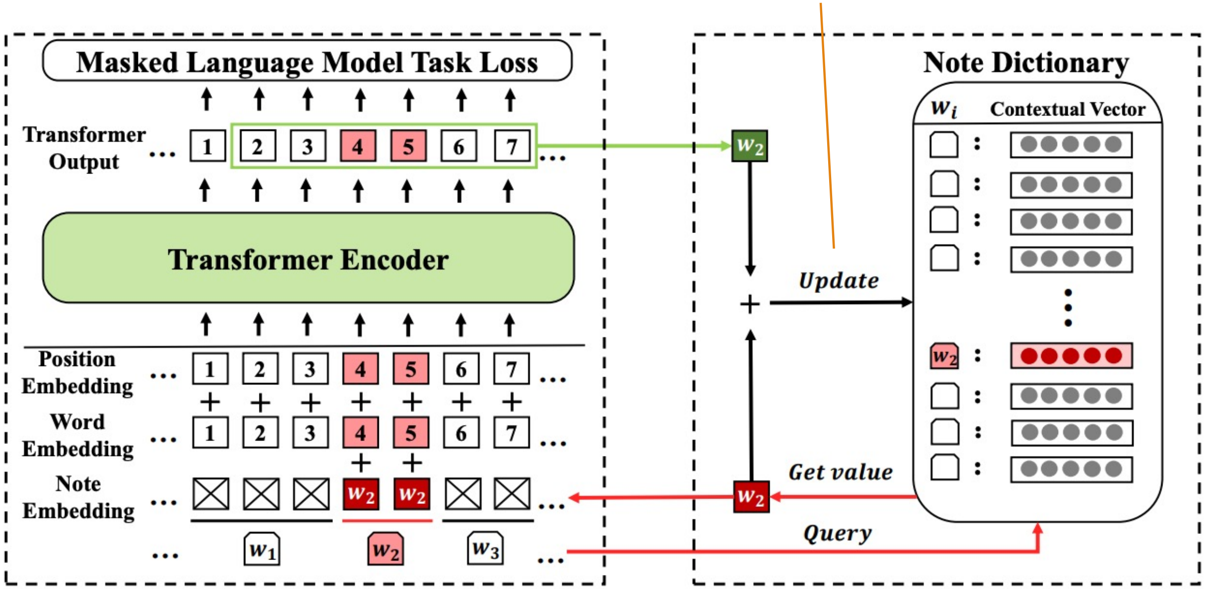
Note-taking is a useful skill which can help people recall information that would otherwise be lost.



In our dataset (Wikipedia and BookCorpus containing 3.47B words), 20% of sentences and 90% of inputs contain at least one rare word (200K with frequency 100 - 500).

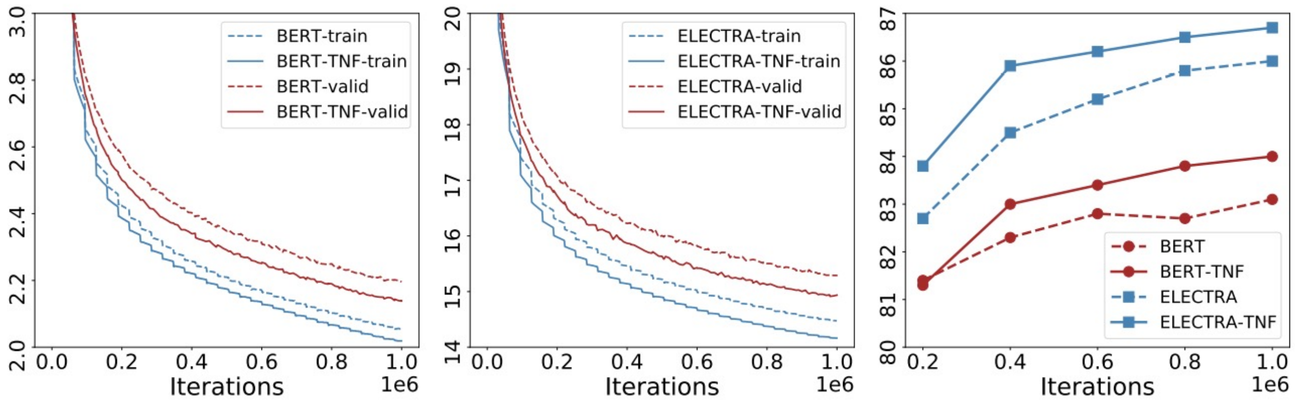
Taking notes helps language pre-training

Exponential moving average



W2 is a rare word

Taking notes expedites language pre-training



(a) Loss curves (BERT setting) (b) Loss curves (ELECTRA setting) (c) GLUE evaluation

Figure 3: The curves of pre-training loss, pre-training validation loss and average GLUE score for all models trained under the BERT setting and ELECTRA setting. All three sub-figures show that TNF expedites the backbone methods.

Save 60% pretraining time!

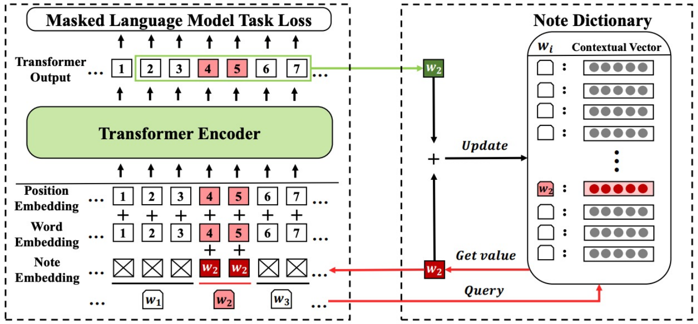
Note dictionary can be removed after pre-training is finished

	MNLI	QNLI	QQP	SST	CoLA	MRPC	RTE	STS	Avg.
BERT (Ours)	85.0	**91.5**	91.2	93.3	58.3	88.3	69.0	88.5	83.1
BERT-TNF	85.0	91.0	**91.2**	93.2	59.5	**89.3**	**73.2**	**88.5**	**83.9**
BERT-TNF-F	**85.1**	90.8	91.1	93.3	59.8	88.8	72.1	88.5	83.7
BERT-TNF-U	85.0	90.9	91.1	**93.4**	**60.2**	88.7	71.4	88.4	83.6
ELECTRA(Ours)	86.8	92.7	91.7	93.2	66.2	**90.2**	76.4	**90.5**	86.0
ELECTRA-TNF	**87.0**	**92.7**	**91.8**	93.6	**67.0**	90.1	81.2	90.1	**86.7**
ELECTRA-TNF-F	86.9	92.6	91.8	**93.7**	65.9	89.7	**81.4**	89.8	86.5
ELECTRA-TNF-U	86.9	92.7	91.7	93.6	66.3	89.8	81.0	89.8	86.5

Table 2: Performance of different models on downstream tasks. Results show that TNF outperforms backbone methods on the majority of individual tasks. We also list the performance of two variants of TNF. Both of them leverage the node dictionary during fine-tuning. Specifically, TNF-F uses fixed note dictionary and TNF-U updates the note dictionary as in pre-training. Both models outperforms the baseline model while perform slightly worse than TNF.

Takeaways

1. Rare words make input noisy, which can slow down optimization of the whole model.
2. Taking notes during the pre-training can outperform baselines on GLUE with 40% pre-training time.
3. The note dictionary can be removed after the pre-training is finished.



Utilize cross-sentence signal to address rare words issue in language model pre-training(Wu et al. ICLR 2021)

Beyond Models: Data-Centric Approaches

Large-scale unlabeled text produce a wealth of information, but still imperfect.

My research focus lies in **better utilization of unlabeled text to improve NLP models**

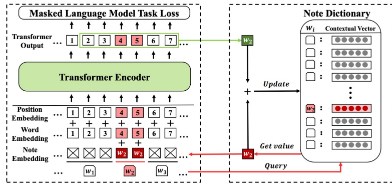
Dataset construction from
- Collection
- Input format

Dataset augmentation

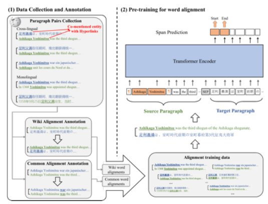
Data quality

…

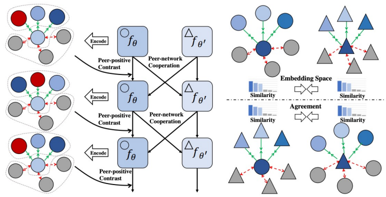
Utilize cross-sentence signal to address rare words issue in language model pre-training(Wu et al. ICLR 2021)



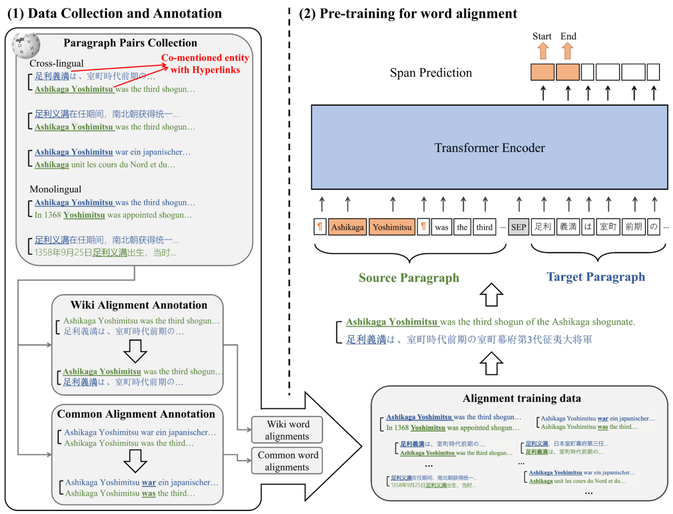
Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training (Wu et al. ACL 2023)



Unsupervised data augmentation for sentence embedding by contrastive learning(Wu et al. EMNLP 2022)



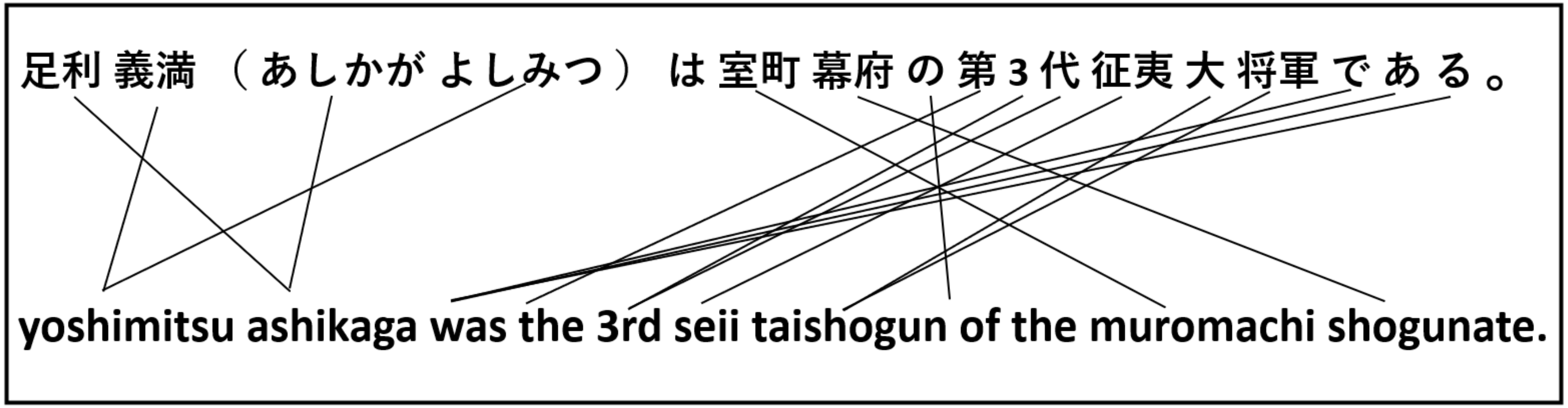
Beyond Models: Data-Centric Approaches



Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training (Wu et al. ACL 2023)

Word Alignment

word alignment aims to align the corresponding words in parallel texts.



Do we really need manual alignment data to do word alignment?

Most existing word alignment methods rely on either **manual alignment datasets or parallel corpora** for training, which weakens their usefulness because of the limiting accessibility of data.

We relax the requirements for:

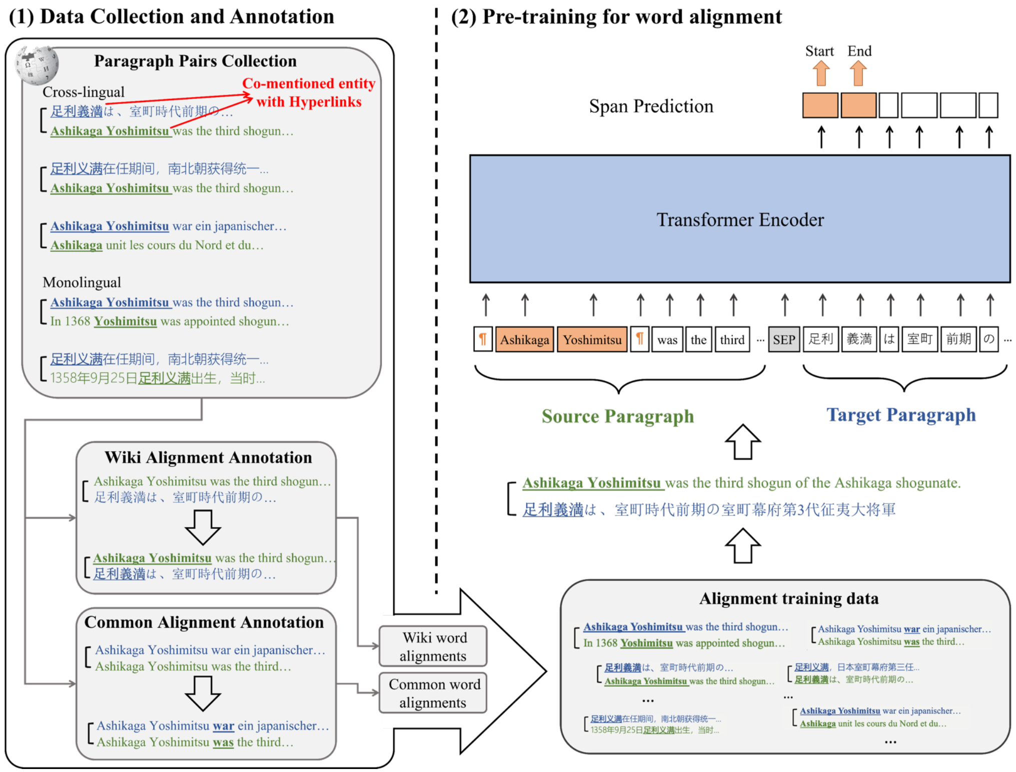
- **correct** (manually made),
- **fully-aligned** (all words in a sentence pair are annotated),
- **parallel sentences**.

Specifically, we make **a large-scale (2 million pairs) training data** that are:

- **noisy (automatically made),**
- **partially-aligned**,
- **non-parallel paragraphs** (or mono-lingual paragraph pairs).

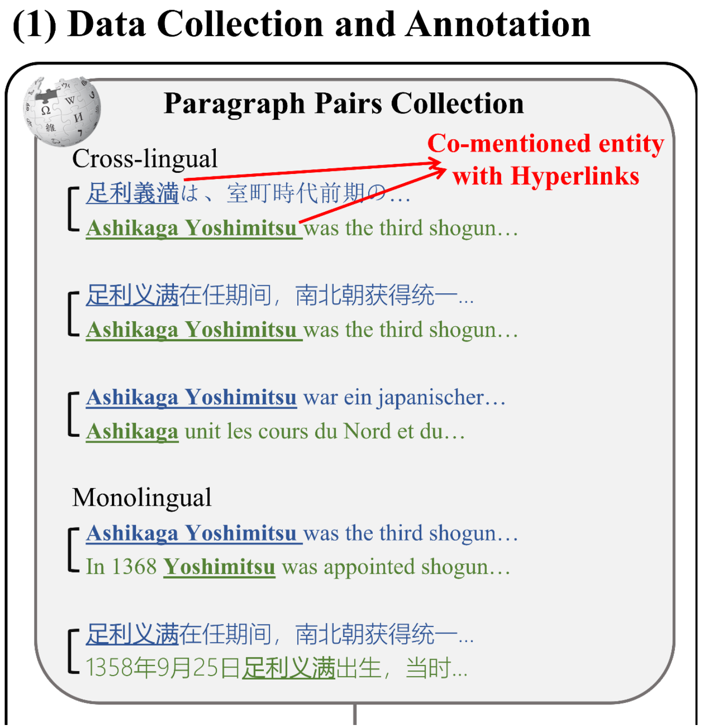
Approach: word alignment pre-training via large-scale weakly supervised span prediction

- **Data Collection**
- **Common word annotation**
- **Wiki word Annotation**
- **Span-prediction Pre-training**



Paragraph pair collection

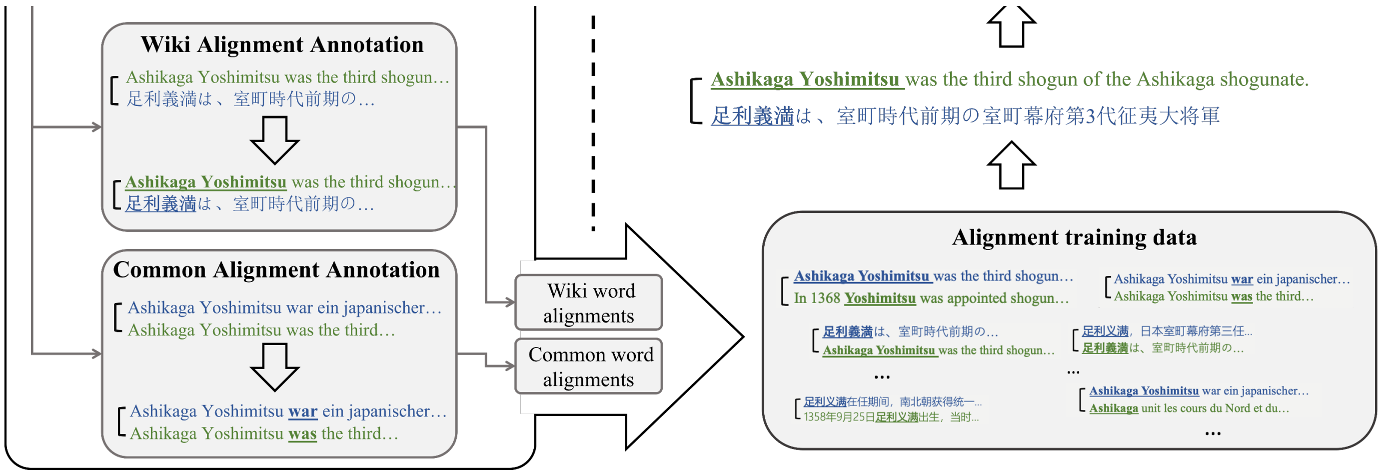
- **Data Collection**
- Common word annotation
- Wiki word Annotation
- Span-prediction Pre-training



Collect both mono-lingual and Cross-lingual Wikipedia paragraph pairs by **co-mentioned hyperlinks**.

Alignment annotation

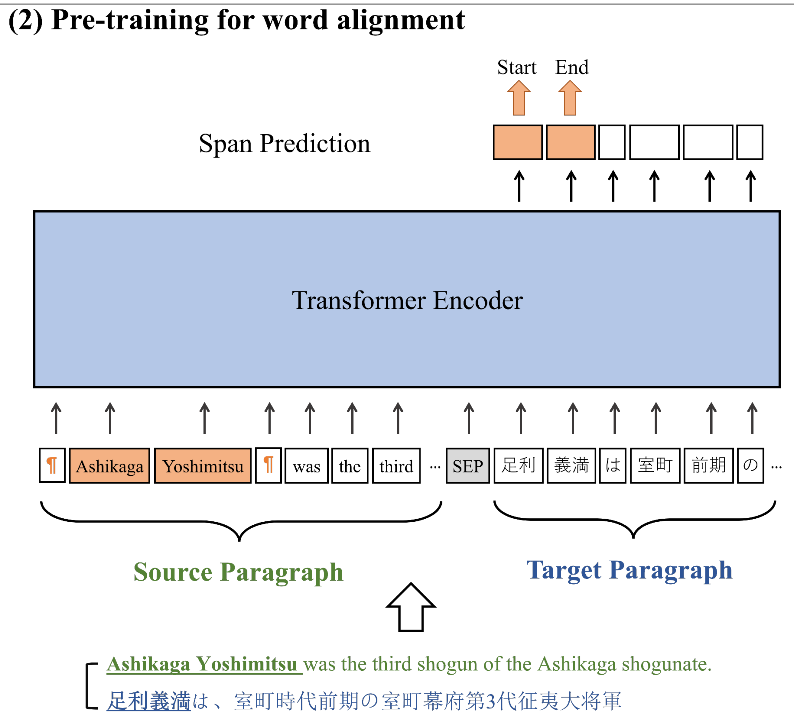
- Data Collection
- **Common word annotation**
- **Wiki word Annotation**
- Span-prediction Pre-training



- **Make Common word annotation** by bi-directional agreement, with **contextual embeddings** in a pre-trained language model.
- **Make Wiki word Annotation** by directly aligning the corresponding **hyperlinks spans of the co-mentioned entity**.

Span prediction pre-training

- ➢ Data Collection
- ➢ Common word annotation
- ➢ Wiki word Annotation
- ➢ **Span-prediction Pre-training**



- ○ Given a source paragraph with a source token specified by the **special token** ¶, the goal is to predict the aligned tokens in the target paragraph.
- ○ Concatenate the source and target paragraph as input sequence and perform the **span prediction task**.

Experiments

Test Set	Method	Precision	Recall	F1	AER
Zh-En	FastAlign (Stengel-Eskin et al.)	80.5	50.5	62.0	-
	DiscAlign (Stengel-Eskin et al.)	72.9	74.0	73.4	-
	SpanAlign (Nagata et al., 2020)	84.4	89.2	86.7	13.3
	WSPAlign (ours)	90.8	92.2	**91.5** (↑ 4.8)	**8.5** (↓ 4.8)
Ja-En	Giza++ (Neubig, 2011)	59.5	55.6	57.6	42.4
	AWESoME (Dou and Neubig, 2021)	-	-	-	37.4
	SpanAlign (Nagata et al., 2020)	77.3	78.0	77.6	22.4
	WSPAlign (ours)	81.6	85.9	**83.7** (↑ 6.1)	**16.3** (↓ 6.1)
De-En	SimAlign (Jalili Sabet et al., 2020)	-	-	81.0	19.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	15.0
	SpanAlign (Nagata et al., 2020)	89.9	81.7	85.6	14.4
	WSPAlign (ours)	90.7	87.1	**88.9** (↑ 3.3)	**11.1** (↓ 3.3)
Ro-En	SimAlign (Jalili Sabet et al., 2020)	-	-	71.0	29.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	20.8
	SpanAlign (Nagata et al., 2020)	90.4	85.3	86.7	12.2
	WSPAlign (ours)	92.0	90.9	**91.4** (↑ 4.7)	**8.6** (↓ 3.6)
En-Fr	SimAlign (Jalili Sabet et al., 2020)	-	-	93.0	7.0
	AWESoME (Dou and Neubig, 2021)	-	-	-	4.1
	SpanAlign (Nagata et al., 2020)	97.7	93.9	-	4.0
	WSPAlign (ours)	98.8	96.0	-	**2.5** (↓ 1.5)

Table 1: Comparison of WSPAlign and previous methods on word alignment datasets. Higher F1 scores are better. Lower AER scores are better. We highlight the best number in the same setting and test set with bold font.

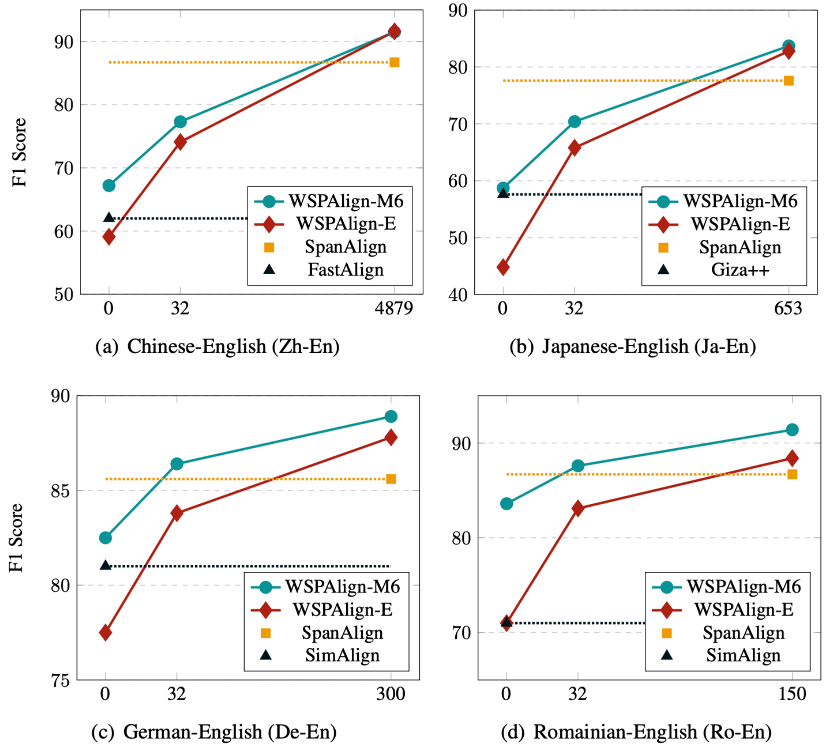
Few-shot, fine-tuning and mono-lingual pre-training

Multilingual Pre-training

Monolingual Pre-training

Unsupervised Baseline

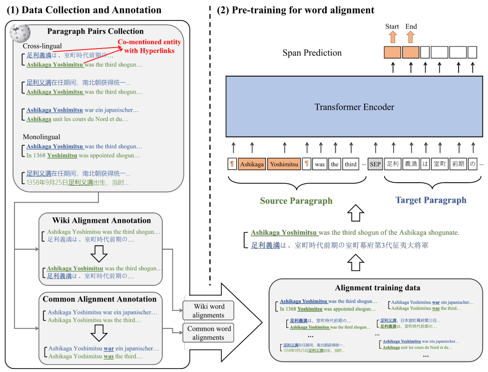
Supervised Baseline



- WSPAlign can be significantly improved and outperforms the existing unsupervised baselines with **few-shot** examples, which can be collected at a low cost.
- If we further fine-tune WSPAlign with a **full supervised dataset**, it can outperform the supervised baseline on all test sets.
- The improvement holds for **mono-lingual pre-training**.

Takeaways

1. We don't have to make perfect (correct, fully-aligned, parallel corpus) datasets to train word aligner.
2. Instead, weak supervision in large-scaled unlabeled text (noisy, partial, non-parallel) can be utilized for pre-training.
3. Zero-shot WSPAlign can outperform unsupervised baseline; few-shot and full-shot finetuning can further improve it and outperform supervised baseline.
4. Mono-lingual pre-training can be transferred to cross-lingual evaluation.



Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training (Wu et al. ACL 2023)

Beyond Models: Data-Centric Approaches

Large-scale unlabeled text produce a wealth of information, but still imperfect.

My research focus lies in **better utilization of unlabeled text to improve NLP models**

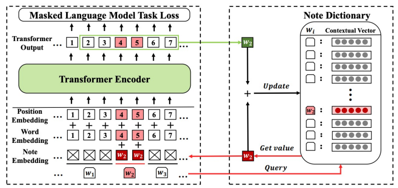
Dataset construction from
- Collection
- Input format

Dataset augmentation

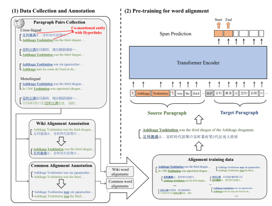
Data quality

…

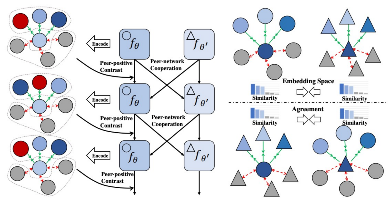
Utilize cross-sentence signal to address rare words issue in language model pre-training(Wu et al. ICLR 2021)



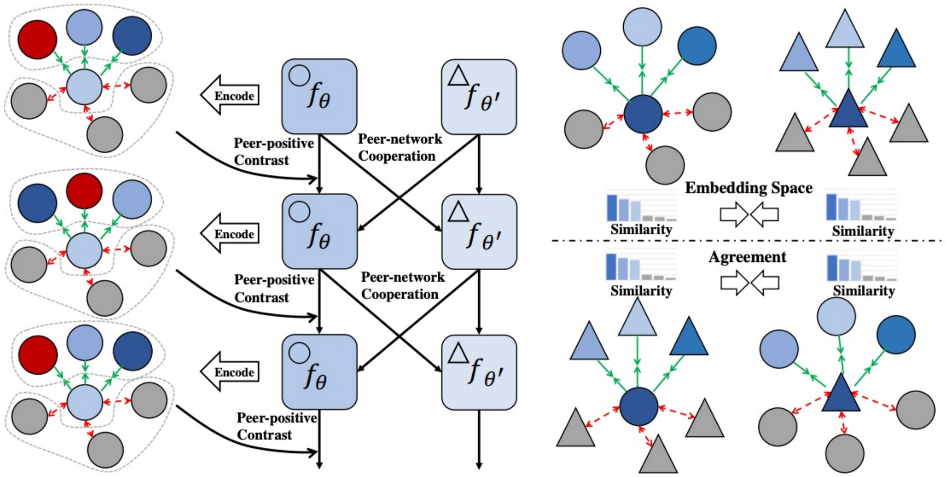
Utilize co-mentioned entities to construct weakly-supervised for word alignment pre-training (Wu et al. ACL 2023)



Unsupervised data augmentation for sentence embedding by contrastive learning(Wu et al. EMNLP 2022)



Beyond Models: Data-Centric Approaches



Unsupervised data augmentation for sentence embedding by contrastive learning(Wu et al. EMNLP 2022)

Biases in unsupervised sentence embedding with contrastive learning

Contrastive learning is a common solution for sentence embedding

Data augmentation can construct positive instance.

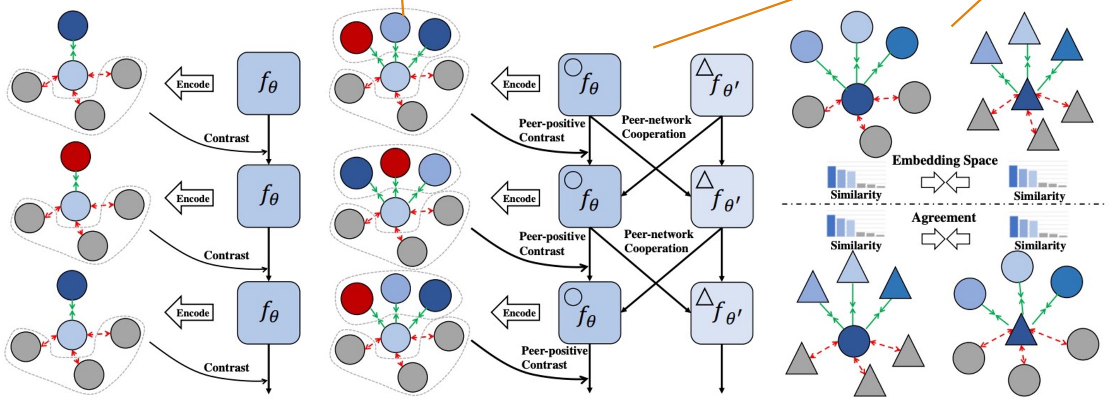
However, text augmentation strategies change semantics in the sentence but still has shortcuts to learn.

Augmenting	**Order**	**N-gram**	**Bag-of-words**
Shuffled Sentence	×	×	✓
Inversed Sentence	×	✓	✓
Word Repetition	✓	×	✓
Word Deletion	✓	×	×

Utilize multiple and diverse augmentations to construct training examples

Multiple augmentations introduce more noises, peer-network makes training more robust

At least one effective positive to learn



(a) Conventional CL (b) PCL with diverse augmentations (c) Cooperative learning objective

Both the number and diversity of augmentations are important

More augmentations is better!

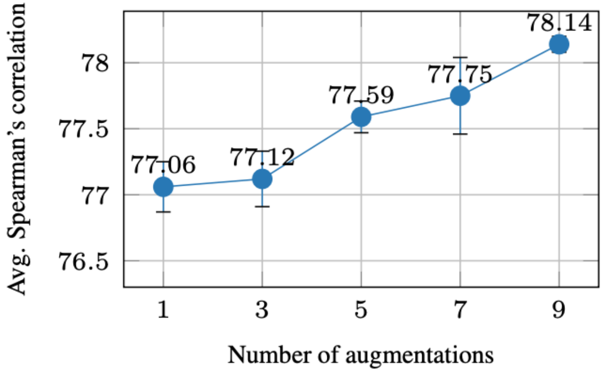


Figure 2: Effect of the number of augmentations.

Diverse augmentation is better!

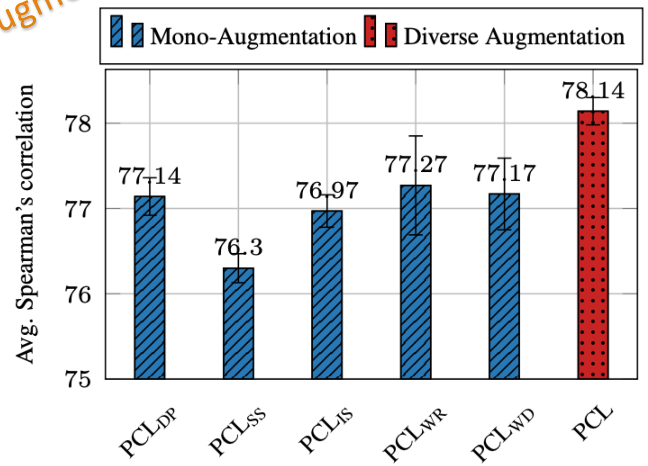
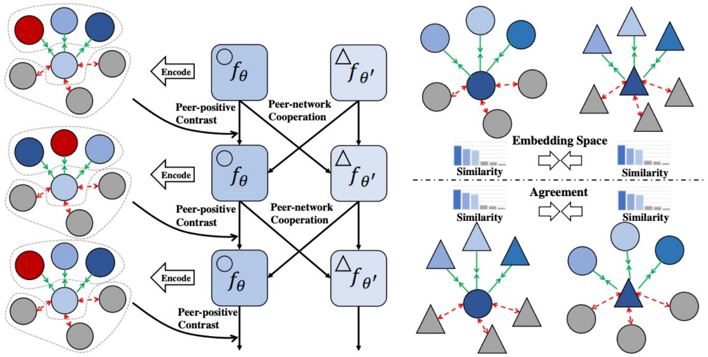


Figure 3: Effect of the diversity of augmentations.

Takeaways

1. Single augmentation can be biased in contrastive learning for unsupervised sentence embeddings.
2. Utilizing multiple and diverse augmentation can mitigate the bias issue.
3. Dual networks can make the training with multiple positives more robust.



Unsupervised data augmentation for sentence embedding by contrastive learning(Wu et al. EMNLP 2022)

Thank you for your attention!

ANY COMMENTS ARE WELCOMED

QIYU WU

CONTACT: WUQIYU576@GMAIL.COM