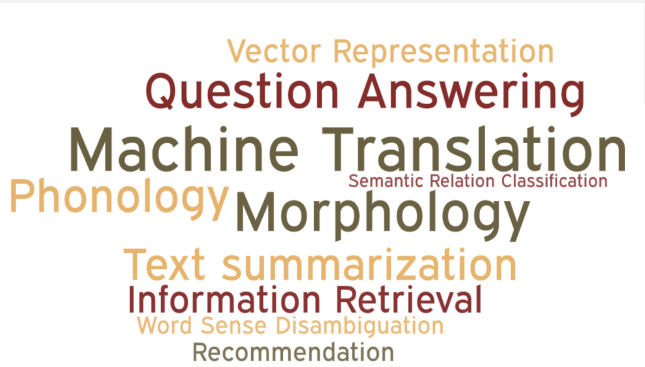
Masked prompt learning for formal analogies beyond words

Liyan Wang Yves Lepage

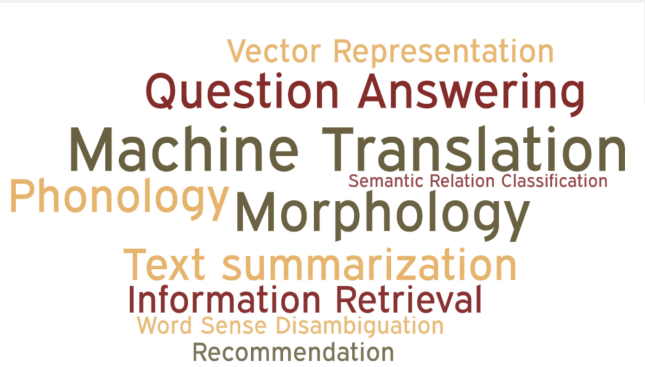
EBMT / NLP Laboratory
IPS, Waseda University

Analogy beyond words



- **Applications:**
  - machine translation (Lepage and Denoual, 2005)
  - question answering (Diallo et al., 2019)
  - text summarization (Elayeb et al., 2020)
- **Challenges:**
  - linguistic complexity
  - the vector offsets may **not** be kept within sentence embeddings
  - extractive mechanism requires predefined **candidate** answers

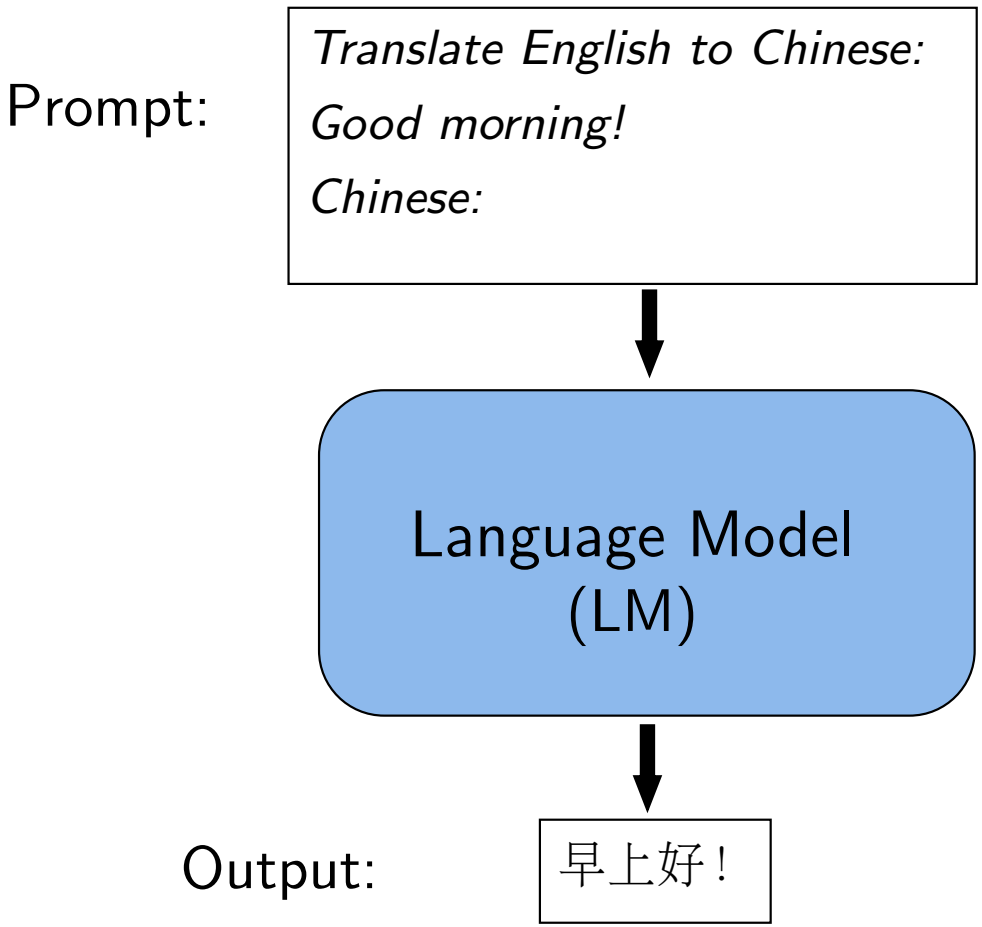
Analogy beyond words



- **Applications:**
  - machine translation (Lepage and Denoual, 2005)
  - question answering (Diallo et al., 2019)
  - text summarization (Elayeb et al., 2020)
- **Challenges:**
  - linguistic complexity
  - the vector offsets may **not** be kept within sentence embeddings
  - extractive mechanism requires predefined **candidate** answers

Our goal is to derive a generative model with the ability of solving analogies beyond words.

Prompt learning



Template

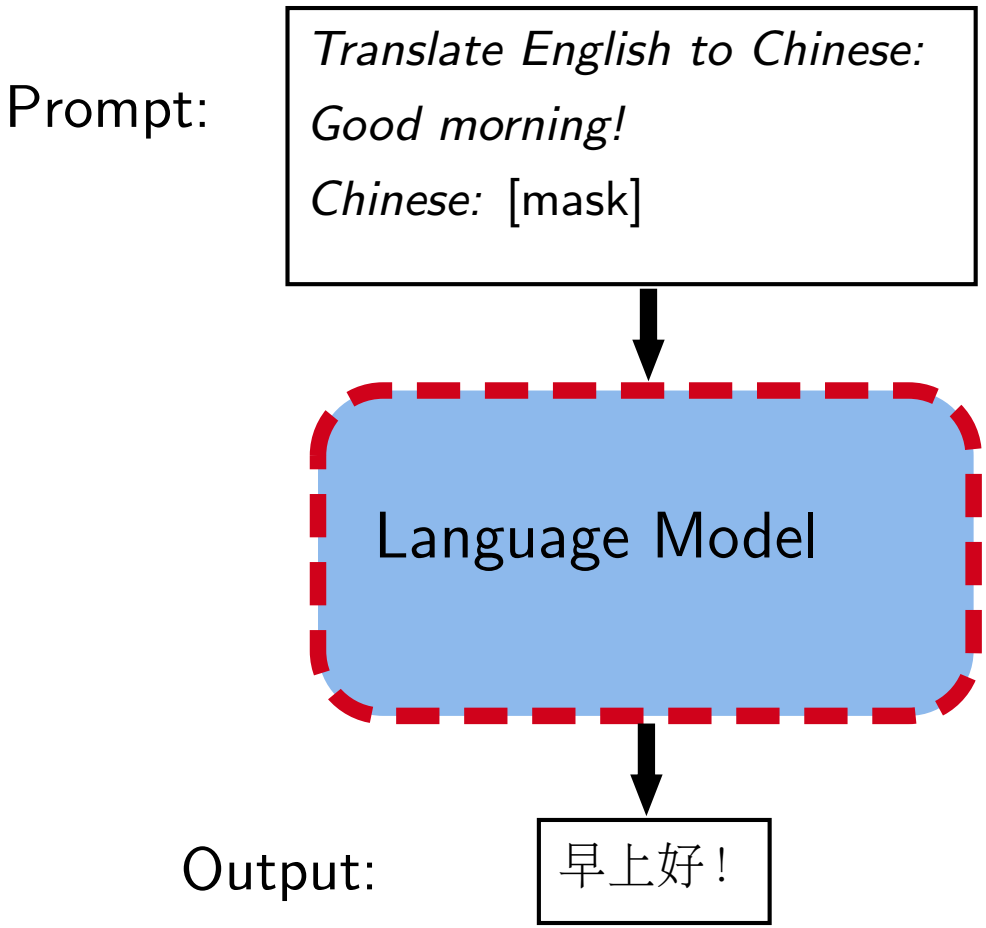
[task description]

[query text]

[output indicator]

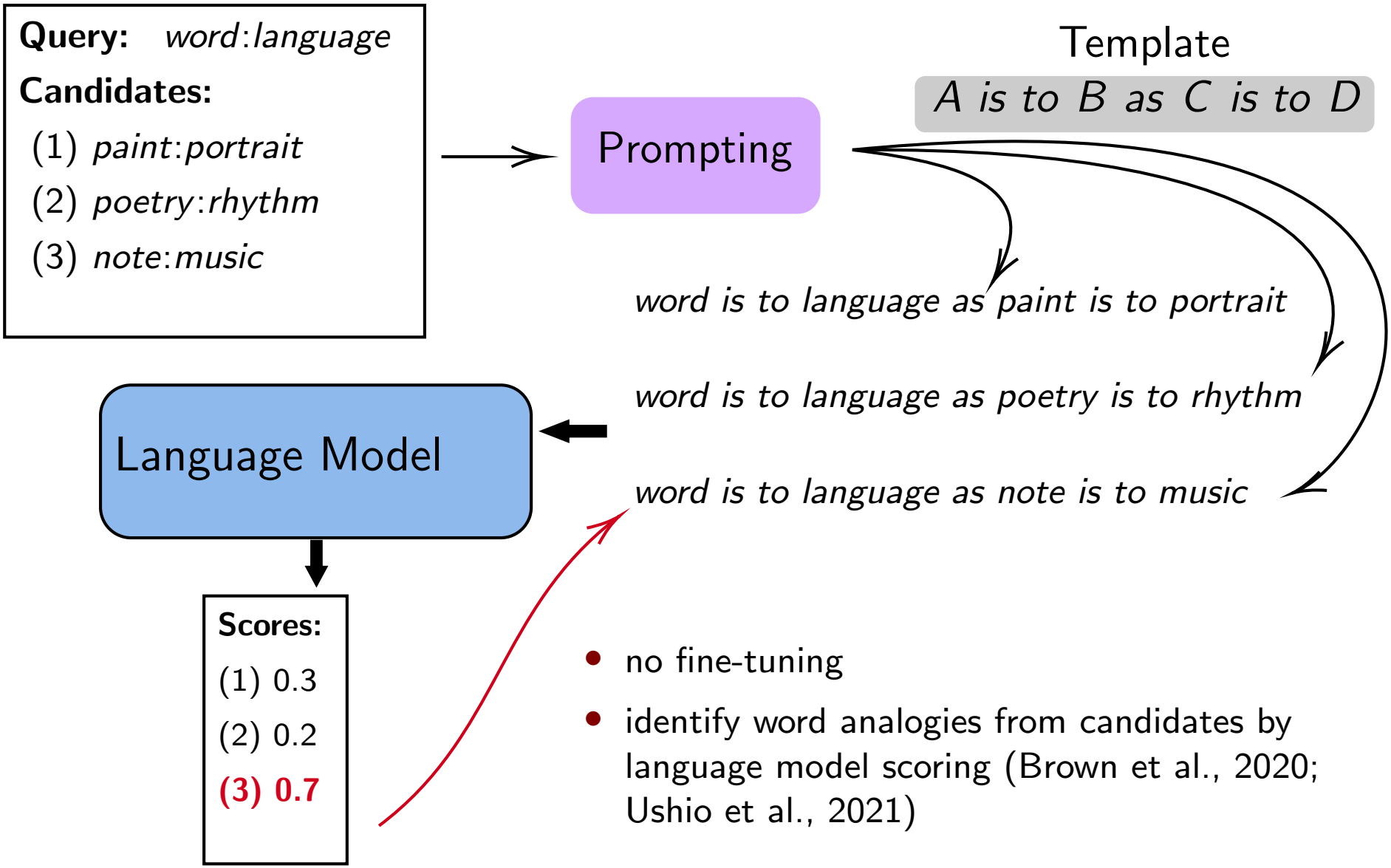
- no fine-tuning
- tasks ≈ cloze-style problem
- input prompt guides the pre-trained LM to generate target output

Prompt-based fine-tuning



- gradient updates
- tasks $\approx$ masked language modeling on prompts
- $[\mathrm{mask}] =$ target outputs of downstream tasks

Prompt learning on analogy



- no fine-tuning
- identify word analogies from candidates by language model scoring (Brown et al., 2020; Ushio et al., 2021)

This work

- **Prompt learning**

	fine-tuned	target-agnostic
prompt learning	✗	-
prompt-baesd fine-tuning	✓	✗
this work	✓	✓

- **Analogy**

	fine-tuned	generative	unit
(Brown et al., 2020; Ushio et al., 2021)	✗	✗	word
this work	✓	✓	phrase, sentence

Motivation

$$A : B :: C : x$$

Motivation

$$A : B :: C : x$$

$$A : B :: C : [\text{mask}]$$

Can we formulate the analogy task as masked analogy completion problem?

Motivation

$$A : B :: C : x$$

$$A : B :: C : [\text{mask}]$$

Can we formulate the analogy task as masked analogy completion problem?

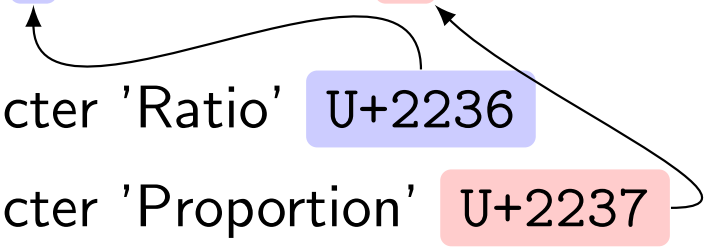
- prompting
- masking strategy
- learning procedure

Symbolic prompt template

Analogical quadruples are converted into sequences.

Example:

he will come tomorrow. : *he will come.* :: *i have no time tomorrow.* : *i have no time.*



- Unicode Character 'Ratio' `U+2236`
- Unicode Character 'Proportion' `U+2237`

Masking schemes

- Target-oriented masking (`[mask]` = **term** D)
- One-term masking (`[mask]` = **any term**)
- Arbitrary masking (`[mask]` = **any span**)

Example:

he will come tomorrow. : *he will come.* :: *i have no time tomorrow.* : *i have no time.*

`[mask]` = '*i have no time.*'

◂ Appendix

Masking schemes

- Target-oriented masking (`[mask]` = **term** D)
- One-term masking (`[mask]` = **any term**)
- Arbitrary masking (`[mask]` = **any span**)

Example:

he will come tomorrow. : *he will come.* :: *i have no time tomorrow.* : *i have no time.*

`[mask]` = '*i have no time tomorrow.*'

◂ Appendix

Masking schemes

- Target-oriented masking (`[mask]` = **term** D)
- One-term masking (`[mask]` = **any term**)
- Arbitrary masking (`[mask]` = **any span**)

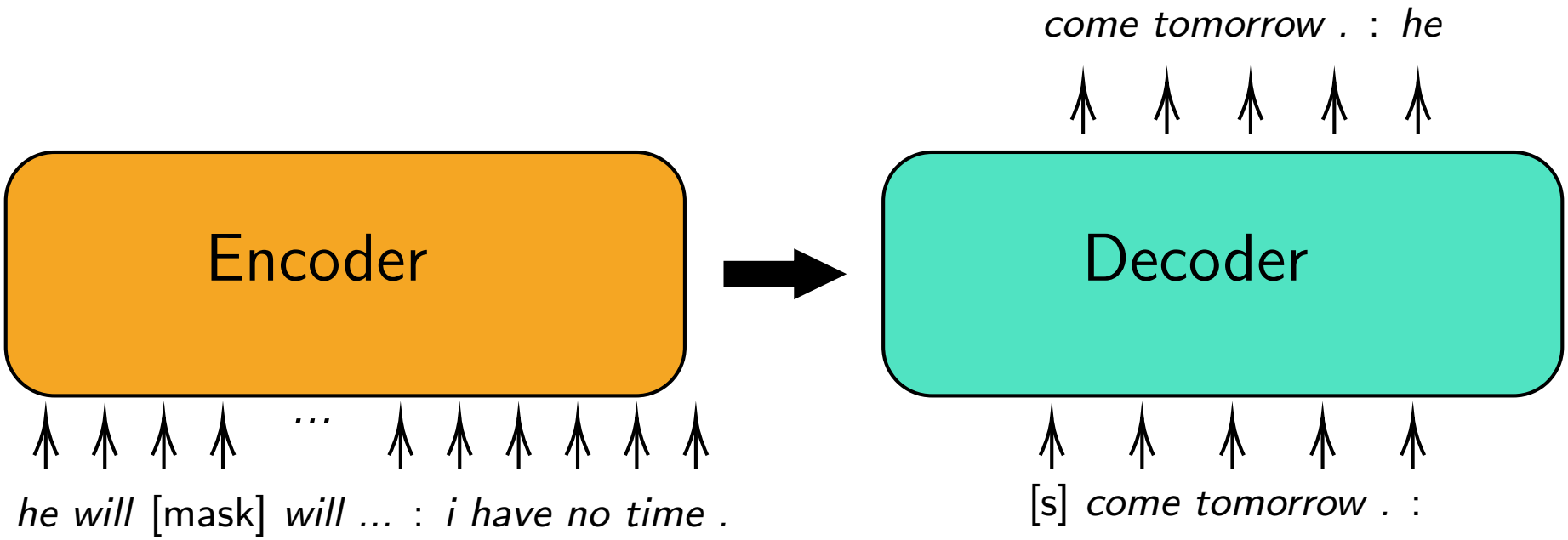
Example:

he will come tomorrow. : he will come. :: *i have no time tomorrow. : i have no time.*

`[mask]` = '*come tomorrow. : he*'

◂ Appendix

Learning paradigm



- **Input** X : the sequence of a masked prompt
- **Output** Y: the sequence of the masking span
- **Objective**: to maximize the conditional probability

$$P(Y|X) = \prod_{t=1}^{|Y|} P(Y_t | Y_{<t}, X; \Theta) \quad (1)$$

Data used for experiments

- **Phrase analogies:**
  - `Berkeley Neural Parser`[1] to extract constituencies (i.e., phrases) between 2 and 6 in length
  - functions from the `Nlg` tools[2] to extract formal analogies
    - representing each phrase as a bag-of-word vector using `Nlg.Lines2Vectors`
    - running `Nlg.Vectors2Clusters` to find analogical clusters
    - any two ratios in a cluster makes an analogy
  - 3,003 English sentences from WMT20[3]
- **Sentence analogies:**
  - the English part of the bilingual analogies used in example-based machine translation taken from (Taillandier et al., 2020)

◂ Appendix

[1] `https://github.com/nikitakit/self-attentive-parser`
[2] `http://lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-15k00317/`
[3] `http://www.statmt.org/wmt20/translation-task.html`

Statistics

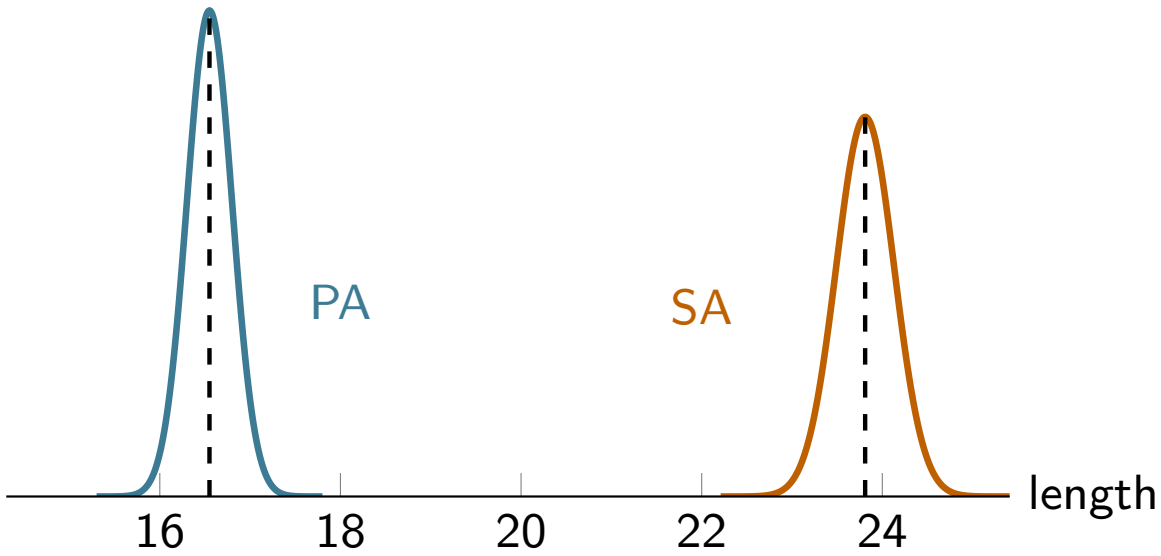


Figure: Distribution of sequence lengths of answered prompts in two analogy data sets: Phrase Analogies (PA) and Sentence Analogies (SA).

	PA (In-distributed)	SA (Out-of-distributed)
Training	1,500,000	0
Validation	1,000	0
Test	1,000	1,000

Training details

- initializing with a pre-trained **BART** model
  - `bart-base`[4]
  - a 6-layer encoder + a 6-layer decoder
- conducting **partial freezing** on the model
  - frozen layers: `encoder[2:]` + `decoder[:4]`
  - ⅓ of layers fine-tuned
- applying **early stopping** with the patience of 2 epochs

[4] `https://huggingface.co/facebook/bart-base`

Comparising models

- autoregressive LM (**GPT-2**[5]) fine-tuned on prompts
- sequence-to-sequence LM (**BART**) with different masking schemes
  - using fine-tuning strategies of tuned layers and prompt templates with ablation

[5] `https://huggingface.co/distilgpt2`

Data	Model	Masking scheme	Acc (%)	distance in chars
PA	GPT-2	-	**99.7**	**0.01±0.01**
	BART	Term *D*	50.9	0.58±0.07
		Any term	97.0	0.05±0.03
		Any span	97.5	0.05±0.03
SA	GPT-2	-	4.2	12.92±0.83
	BART	Term *D*	12.0	3.17±0.30
		Any term	**44.4**	**2.85±0.29**
		Any span	11.4	4.28±0.38

Table: Comparison between the autoregressive baseline and masked prompt learning with different masking schemes over two analogy test sets.

- **Model architecture:**
  - BART $\lessapprox$ GPT-2

Data	Model	Masking scheme	Acc (%)	distance in chars
PA	GPT-2	-	**99.7**	**0.01±0.01**
	BART	Term D	50.9	0.58±0.07
		Any term	97.0	0.05±0.03
		Any span	97.5	0.05±0.03
SA	GPT-2	-	4.2	12.92±0.83
	BART	Term D	12.0	3.17±0.30
		Any term	**44.4**	**2.85±0.29**
		Any span	11.4	4.28±0.38

Table: Comparison between the autoregressive baseline and masked prompt learning with different masking schemes over two analogy test sets.

- **Masking scheme:**
  - Term $D <$ Any term $\lessapprox$ Any span

Data	Model	Masking scheme	Acc (%)	distance in chars
PA	GPT-2	-	**99.7**	**0.01±0.01**
	BART	Term *D*	50.9	0.58±0.07
		Any term	97.0	0.05±0.03
		Any span	97.5	0.05±0.03
SA	GPT-2	-	4.2	12.92±0.83
	BART	Term *D*	12.0	3.17±0.30
		Any term	**44.4**	**2.85±0.29**
		Any span	11.4	4.28±0.38

Table: Comparison between the autoregressive baseline and masked prompt learning with different masking schemes over two analogy test sets.

- **Out-of-distribution generalization:**
  - GPT-2 $<$ BART
  - Any span $\lessapprox$ Term *D* $<$ Any term

Takeaways

- GPT-2 may overfit to the narrow distribution of phrase analogies, which leads to a failure in solving analogies between longer sequences.
- BART has competitive performance on phrase analogies and exhibits better out-of-distribution generalization.
- Masked prompt learning with one-term masking has the best generalization capability.

Input	*he is getting better.* : *tom is getting better.* :: *he doesn 't like to lose.* : [mask]
Output	*tom doesn 't like to lose*
Reference	*tom doesn 't like to lose .*
Input	*what is going on here?* : *he is intelligent.* :: *what's going on here?* : [mask]
Output	*he i s intelligent.*
Reference	*he ' s intelligent.*
Input	*he will come tomorrow.* : *he will come.* :: *i have no time tomorrow.* : [mask]
Output	*he will have no time tomorrow .*
Reference	*i have no time.*

Table: Examples of analogy results generated by the BART model fine-tuned with one-term masking.

Exploration of analogical ability

$$[\text{mask}] : B :: C : D \qquad \text{Acc}_A$$

$$A : [\text{mask}] :: C : D \qquad \text{Acc}_B$$

$$A : B :: [\text{mask}] : D \qquad \text{Acc}_C$$

$$A : B :: C : [\text{mask}] \qquad \text{Acc}_D$$

$\Big\}$ Acc_{all}

$$\text{Acc}_{\text{all}} = \begin{cases} 1, & \text{if } \text{Acc}_X = 1 \text{ for all } X = A, B, C, D. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Exploration of analogical ability

Data	Masking Scheme	Acc (%) A	B	C	D	all
PA	Term D	0.0	0.0	0.0	50.9	0.0
	Any term	99.8	99.7	**99.7**	97.0	96.5
	Any span	**99.9**	**99.9**	**99.7**	**97.5**	**97.3**
SA	Term D	1.2	0.0	0.0	12.0	0.0
	Any term	**24.9**	**39.3**	**44.0**	**44.4**	**12.9**
	Any span	10.8	10.6	15.7	11.4	0.6

Table: Average accuracy results of fine-tuned BART models with different masking schemes in solving analogy questions in different formats.

- Term D < Any span < Any term

◂ Appendix

Exploration of analogical ability

Data	Masking Scheme	Acc (%) A	B	C	D	all
PA	Term D	0.0	0.0	0.0	50.9	0.0
	Any term	99.8	99.7	**99.7**	97.0	96.5
	Any span	**99.9**	**99.9**	**99.7**	**97.5**	**97.3**
SA	Term D	1.2	0.0	0.0	12.0	0.0
	Any term	**24.9**	**39.3**	**44.0**	**44.4**	**12.9**
	Any span	10.8	10.6	15.7	11.4	0.6

Table: Average accuracy results of fine-tuned BART models with different masking schemes in solving analogy questions in different formats.

- Term D < Any span < Any term

◂ Appendix

Exploration of analogical ability

Data	Masking Scheme	Acc (%) A	B	C	D	all
PA	Term D	0.0	0.0	0.0	50.9	0.0
	Any term	99.8	99.7	**99.7**	97.0	96.5
	Any span	**99.9**	**99.9**	**99.7**	**97.5**	**97.3**
SA	Term D	1.2	0.0	0.0	12.0	0.0
	Any term	**24.9**	**39.3**	**44.0**	**44.4**	**12.9**
	Any span	10.8	10.6	15.7	11.4	0.6

Table: Average accuracy results of fine-tuned BART models with different masking schemes in solving analogy questions in different formats.

- Term D < Any span < Any term

◂ Appendix

Fine-tuned layers

Fine-tuned layer		Acc (%)	
Encoder	Decoder	PA	SA
None	None	0.0	0.0
All	All	94.2	11.7
None	All	88.6	23.7
All	None	70.5	28.7
Bottom two	Top two	**97.0**	**44.4**
Top two	Bottom two	95.8	42.0

Table: Impact of fine-tuning strategies of fine-tuned layers in the pre-trained sequence-to-sequence model.

- off-the-shelf LM failed in the zero-shot setting
- full-scale fine-tuning biased towards the training distribution
- lightweight joint fine-tuning on both encoder and decoder

Fine-tuned layers

Fine-tuned layer		Acc (%)	
Encoder	Decoder	PA	SA
None	None	0.0	0.0
All	All	94.2	11.7
None	All	88.6	23.7
All	None	70.5	28.7
Bottom two	Top two	**97.0**	**44.4**
Top two	Bottom two	95.8	42.0

Table: Impact of fine-tuning strategies of fine-tuned layers in the pre-trained sequence-to-sequence model.

- off-the-shelf LM failed in the zero-shot setting
- full-scale fine-tuning biased towards the training distribution
- lightweight joint fine-tuning on both encoder and decoder

Fine-tuned layers

Fine-tuned layer		Acc (%)	
Encoder	Decoder	PA	SA
None	None	0.0	0.0
All	All	94.2	11.7
None	All	88.6	23.7
All	None	70.5	28.7
Bottom two	Top two	**97.0**	**44.4**
Top two	Bottom two	95.8	42.0

Table: Impact of fine-tuning strategies of fine-tuned layers in the pre-trained sequence-to-sequence model.

- off-the-shelf LM failed in the zero-shot setting
- full-scale fine-tuning biased towards the training distribution
- lightweight joint fine-tuning on both encoder and decoder

Prompt templates

Template	Model	Acc (%) PA	SA
$A : B :: C : D$	GPT-2	**99.7**	4.2
	BART	97.0	**44.4**
A is to B as C is to D	GPT-2	**99.9**	3.6
	BART	99.6	**37.9**
A B C D	GPT-2	**99.9**	0.0
	BART	99.6	0.0

Table: Comparison between different prompt templates: symbolic prompt ($A : B :: C : D$), textual prompt (*A is to B as C is to D*), and null prompt (*A B C D*).

- similar in-distribution accuracy
- behave differently on out-of-distribution cases
- null prompt failed to answer SA
- the necessity of clear prompt semantics

◂ Appendix

Prompt templates

Template	Model	Acc (%) PA	SA
$A : B :: C : D$	GPT-2	**99.7**	4.2
	BART	97.0	**44.4**
A is to B as C is to D	GPT-2	**99.9**	3.6
	BART	99.6	**37.9**
A B C D	GPT-2	**99.9**	0.0
	BART	99.6	0.0

Table: Comparison between different prompt templates: symbolic prompt ($A : B :: C : D$), textual prompt (*A is to B as C is to D*), and null prompt (*A B C D*).

- similar in-distribution accuracy
- behave differently on out-of-distribution cases
- null prompt failed to answer SA
- the necessity of clear prompt semantics

◂ Appendix

Prompt templates

Template	Model	Acc (%) PA	Acc (%) SA
$A : B :: C : D$	GPT-2	**99.7**	4.2
	BART	97.0	**44.4**
A is to B as C is to D	GPT-2	**99.9**	3.6
	BART	99.6	**37.9**
A B C D	GPT-2	**99.9**	0.0
	BART	99.6	0.0

Table: Comparison between different prompt templates: symbolic prompt ($A : B :: C : D$), textual prompt (*A is to B as C is to D*), and null prompt (*A B C D*).

- similar in-distribution accuracy
- behave differently on out-of-distribution cases
- null prompt failed to answer SA
- the necessity of clear prompt semantics

◂ Appendix

Prompt templates

Template	Model	Acc (%) PA	Acc (%) SA
$A : B :: C : D$	GPT-2	**99.7**	4.2
	BART	97.0	**44.4**
A is to B as C is to D	GPT-2	**99.9**	3.6
	BART	99.6	**37.9**
A B C D	GPT-2	**99.9**	0.0
	BART	99.6	0.0

Table: Comparison between different prompt templates: symbolic prompt ($A : B :: C : D$), textual prompt (*A is to B as C is to D*), and null prompt (*A B C D*).

- similar in-distribution accuracy
- behave differently on out-of-distribution cases
- null prompt failed to answer SA
- the necessity of clear prompt semantics

◂ Appendix

Conclusion

- introduced a prompt-based fine-tuning paradigm for solving analogies beyond words
  - masked sequence-to-sequence learning on answered prompts with a symbolic template
- proposed three masking patterns
  - target-oriented masking leads to overfit to narrow features
  - one-term masking is effective for adaptation of generative analogy completion
- lightweight fine-tuning shows the potential for learning robust analogical capability

Conclusion

- introduced a prompt-based fine-tuning paradigm for solving analogies beyond words
  - masked sequence-to-sequence learning on answered prompts with a symbolic template
- proposed three masking patterns
  - target-oriented masking leads to overfit to narrow features
  - one-term masking is effective for adaptation of generative analogy completion
- lightweight fine-tuning shows the potential for learning robust analogical capability

Conclusion

- introduced a prompt-based fine-tuning paradigm for solving analogies beyond words
  - masked sequence-to-sequence learning on answered prompts with a symbolic template
- proposed three masking patterns
  - target-oriented masking leads to overfit to narrow features
  - one-term masking is effective for adaptation of generative analogy completion
- lightweight fine-tuning shows the potential for learning robust analogical capability

Future work

- refine the learning paradigm to enhance out-of-distribution performance in the few-shot scenario
- build a multilingual generator for analogies beyond words

Thank you for your attention!

this point:*any point*	*to say*:*want to say*
this moment:*any moment*	*to go out*:*want to go out*
this legislation:*any legislation*	*be a champion*:*want be a champion*
this idea:*any idea*	
this country:*any country*	
of this country:*of any country*	
this case:*any case*	
at this moment:*at any moment*	

Table: Examples of analogical clusters retrieved from phrases

◂ Content

Masking schemes

- $$\frac{\text{any term} \cap \text{any span}}{\text{any span}} \approx 4\%$$
- $$\frac{\text{term}D \cap \text{any term}}{\text{any term}} \approx 25\%$$

◂ Content

- **Test inputs:**
  - **GPT-2:** $A : B :: C :$
  - **BART:** $A : B :: C : [\mathrm{mask}]$

Input	[mask] : *he teaches english.* :: *she betrayed you.* : *he betrayed you.*
Output	*she teaches english.*
Input	[mask] : *shut the door, please.* :: *close the door.* : *shut the door.*
Output	*close the door, please.*
Input	[mask] : *you're lost.* :: *we're ambitious.* : *you're ambitious.*
Output	*we're lost.*
Input	[mask] : *we could walk.* :: *i can wait.* : *i can walk.*
Output	*we could wait.*
Input	[mask] : *i just don 't want to go with you.* :: *i don 't want to talk to you.* : *i just don 't want to talk to you.*
Output	*i don 't want to go with you.*

Table: Examples of 1.2% sentence analogies that are correctly answered by the model fine-tuned with term *D* masking.

◂ Content

Query	*he's my best friend.* : *you are my best friend.* :: *he's taller than me.* : *x*
Reference	*you are taller than me.*
GPT-2	*his waist meet as retailers meet as retailers*
BART	*you are taller than me*
Query	*you promised.* : *we promised.* :: *you don't have any proof.* : *x*
Reference	*we don't have any proof.*
GPT-2	*we don't have any knowledge of the works*
BART	*we don't have any proof*
Query	*what do you do?* : *i know what to do.* :: *what do you say?* : *x*
Reference	*i know what to say.*
GPT-2	*says much says much*
BART	*to know what*
Query	*i trust you.* : *i don't trust you.* :: *i know his address.* : *x*
Reference	*i don't know his address.*
GPT-2	*Court ruling Byty v. Maxwell v. Maxwell,*
BART	*his address*

Table: Examples of solutions generated by language models fine-tuned on null prompts.

◂ Content

References I

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aïssatou Diallo, Markus Zopf, and Johannes Fürnkranz. 2019. Learning analogy-preserving sentence embeddings for answer selection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 910–919, Hong Kong, China. Association for Computational Linguistics.

Bilel Elayeb, Amina Chouigui, Myriam Bounhas, and Oussama Ben Khiroun. 2020. Automatic Arabic text summarization using analogical proportions. *Cognitive Computation*, 12(5):1043–1069.

Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3):251–282.

Valentin Taillandier, Liyan Wang, and Yves Lepage. 2020. Réseaux de neurones pour la résolution d'analogies entre phrases en traduction automatique par l'exemple (neural networks for the resolution of analogies between sentences in EBMT). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 108–121, Nancy, France. ATALA et AFCP.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.