Continued Pre-training on Sentence Analogies for Translation with Small Data

Liyan Wang, Haotong Wang, Yves Lepage
Waseda University
2-7 Hibikino, Wakamatsu, Kitakyushu, 808-0135, Japan
{wangliyan0905@toki., wanghaotong0925@toki., yves.lepage@}waseda.jp

Abstract

This paper introduces **C**ontinued **P**re-training **o**n **A**nalogies (CPoA) to incorporate pre-trained language models with analogical abilities, aiming at improving performance in low-resource translations without data augmentation. We continue training the models on sentence analogies retrieved from a translation corpus. Considering the sparsity of analogy in corpora, especially in low-resource scenarios, we propose exploring approximate analogies between sentences. We attempt to find sentence analogies that might not conform to formal criteria for entire sentences but partial pieces. When training the models, we introduce a weighting scalar pertaining to the quality of analogies to adjust the influence: emphasizing closer analogies while diminishing the impact of far ones. We evaluate our approach on a low-resource translation task: German-Upper Sorbian. The results show that CPoA using 10 times fewer instances can effectively attain gains of +1.4 and +1.3 BLEU points over the original model in two translation directions. This improvement is more pronounced when there are fewer parallel examples.

Keywords: Sentence Analogies, Continued Pre-training, Analogy Retrieval, Low-resource Translation

1. Introduction and Background

Analogy, succinctly described as A is to B as C is to D, serves as a cognitive shortcut that allows us to generalize, infer, and adapt with exposure to relevant examples (Gentner, 1983; Hofstadter, 2001). This inherent ability stands in stark contrast to the current paradigm in machine learning, where Neural Machine Translation (NMT) models demanding millions of parallel sentences (Fadaee et al., 2017; Haddow et al., 2022) have become normal. Tracing back to Example-Based Machine Translation (Nagao, 1984; Lepage and Denoual, 2005; Taillandier et al., 2020), namely machine translation by analogy, it offers plausible translation answers by reasoning analogies using relatively small data. When translating new sentences, the indirect mechanism (Lepage and Denoual, 2005) efficiently uses parallel examples without the need for additional linguistic knowledge. For instance, to translate the query sentence "*The man loves the woman.*" from English to Chickasaw, the method first retrieves a sentence analogy with the query from the English side. It then infers the translation result by addressing the corresponding analogy in Chickasaw, illustrated as:

The cat chases the dog.	:	*The dog chases the cat.*	::	*The woman loves the man.*	:	***The man loves the woman.***
Kowi'at ofi'ã lhiyohli.	:	*Ofi'at kowi'ã lhiyohli.*	::	*Ihooat hattakã hollo.*	:	***Hattakat ihooã hollo.***

Cognitive translation approaches (Lepage and Denoual, 2005; Langlais, 2013) provide structural representations for sentences in the source and target languages on the basis of formal analogies (Lepage, 1998; Langlais and Yvon, 2008), where the left and right ratios share the same transformations on forms. While translation by analogy can achieve impressive performance with small data, it might struggle when valid analogies cannot be found. This is substantially limited by the closeness between example sentences and queries. Additionally, finding analogies in semantics remains challenging (Gladkova et al., 2016).

In NMT, multilingual language models (Conneau and Lample, 2019; Liu et al., 2020) pre-trained on massive amounts of data have shown to adapt the rich linguistic knowledge to the specifics of translation between source and target languages. The prevailing notion is that more parallel data for fine-tuning leads to better translation performance. However, compiling and enlarging high-quality parallel corpora is arduous, especially for low-resource languages. This leads us to ask: can analogy help pre-trained language models for translation with limited data? Will the intricate knowledge structured in analogies enable the models to grasp the nuances of the languages defined in available data?

Recent research has investigated zero- or few-shot solutions of large language models, such as a series of GPT models, to analogy questions (Ushio et al., 2021; Webb et al., 2023). It becomes increasingly difficult when analogies go beyond words (Wijesiriwardene et al., 2023). By further training using analogy instances, the models exhibit a strong adaptation with explicit improvements in reasoning

LREC-COLING 2024, pages 3890–3896
20-25 May, 2024.

analogies (Wang and Lepage, 2022).

In this work, we introduce a mechanism for continued pre-training using sentence analogies, aiming to adapt pre-trained models to languages in low-resource translation tasks. Concretely, it consists of finding analogies between sentences in the given translation corpus, and update a pre-trained model to capture analogical regularities in approximate scenarios. For translation, we then fine-tune the learnt model on the parallel corpus. We evaluate our approach using a pre-trained mBART model (Liu et al., 2020) on the German-Upper Sorbian language pair. Our approach shows advantages in both translation directions, and even more so in the case of only 10,000 parallel sentences. The retrieved analogies effectively encapsulate the underlying relationships in the limited corpus. With adaptive pre-training, the model is able to learn analogical abilities, attaining better performance in low-resource translations.

2. Approach

From a given translation corpus, our approach attempts to retrieve sentence analogies for the source and target languages independently (Section 2.1). We continue training a pre-trained model on the retrieved data to incorporate the ability of solving analogies in both languages (Section 2.2). For translation, the trained model is then fine-tuned with the parallel data in a specific direction.

2.1. Analogy Retrieval

Retrieval is performed on each side of the parallel data. To recognize general analogical relationships, our approach starts with gathering sentence patterns that preserve the structural form of sentences while abstracting away specific details. By discovering formal analogies among these patterns, we can find approximate analogies by locating sentences that contain analogical patterns. The retrieval procedure is elaborated further below.

Sentence Patterns We employ a syntax parser to analyse the constituents within each sentence. For each constituent, we substitute it with a placeholder[1], represented as "...". The resulting sequence forms a pattern. Typically, a sentence with n constituents will yield n distinct patterns. In this work, we use patterns that retain at least a quarter of the original sentence.

Pattern Analogies We then apply the Nlg package (Fam and Lepage, 2018) to find formal analogies between patterns at the character level. Nlg produces analogical clusters, each of which groups pairs of pattens that have consistent differences in form. In general, in each cluster, every two pairs of patterns form a formal analogy. We consider pattern analogies where parallel transformations are on content pieces other than placeholders.

Approximate Analogies between Sentences We next proceed to construct sentence quadruples drawing from the patterns in the collected analogies. From a formal perspective, these quadruples represent approximate analogies that capture analogical relations through partial sentence content. To measure the quality of a sentence analogy, i.e., the extent of formal analogy present in a sentence quadruple, we compute analogical coverage, defined as the length coverage of sentences by their respective patterns that form a formal analogy, as

$$\mathrm{cov} = \frac{1}{4} \sum_{x \in A,B,C,D} \frac{|\mathrm{excl}(p^x)|}{|q^x|} \tag{1}$$

Here, q refers to a sentence analogy, drawing from a pattern analogy p. q^x represents the term x in q. The function $\mathrm{excl}(p^x)$ is used to eliminate placeholders in the pattern p^x. We compute sequence lengths in characters. In the collection of sentence quadruples, a single quadruple can possess multiple coverage scores. These scores result from different formal analogies between patterns that vary in sentence hierarchy. For each unique approximate analogy, we attach the information of the pattern analogy that contributes to the maximum analogical coverage.

Below is an example of an approximate analogy between English sentences

I want to drive. : *I do not want to drive.* :: *I hope to see you again* : *I do not want to see you again.*

drawn from the pattern analogy

I want to drive. : *I do not want to drive.* :: *I ... to see you again.* : *I do not ... to see you again.*

where about 93% of the content (notice that the expected "*hope*" is substituted by "*want*" in this example) within these sentences captures an analogical relationship at the formal aspect.

2.2. Analogy-adaptive Training

Task Let $\mathcal{Q}$ be a collection of analogies from sentence sets in the source and target languages. We train a language model to solve sentence analogies by generating solutions given three known terms.

[1]We use the Unicode character `u+23EF` as a special token for the placeholder in patterns.

Learning Paradigm As in (Wang and Lepage, 2022), we perform masked sequence-to-sequence learning on masked analogies to reconstruct masking terms (i.e., the unknown sentence in an analogy question). For each quadruple (A, B, C, D), we concatenate the four terms into a consecutive sequence in a specific template S as $A : B :: C : D$`<LID>`, where we use the Unicode characters `U+2236` and `U+2237` as functional tokens for the ratio (:) and the proportion (::). `<LID>` is a language token. We apply the one-term masking scheme on sequences that randomly masks one of the four terms, and train the model to generate the missing term in masked analogies. We optimize the pre-trained model to minimize the divergence between generated solutions and expected answers.

Objective To generalize to approximate analogies, we introduce a weighting scalar that adjusts the model's penalization for different instances based on the quality of analogies. This weighting scalar helps differentiate the impact of instances during training. It allows the model to prioritize instances with higher coverage and avoid bias towards less informative analogies. We train the model for convergence on the weighted cross-entropy loss as:

$$\mathcal{L} = -\sum_{q_i \in \mathcal{Q}} w_i \log P(M_i | S(q_i) \backslash M_i) \tag{2}$$

, where w_i is the analogical coverage of the sentence analogy q_i . $S(q_i) \backslash M_i$ and M_i are the masked sequence and the sequence of the masking term.

3. Experiments

3.1. Setup

Data We experiment with the German-Upper Sorbian (de-hsb) language pair, from the low-resource translation task[2] at WMT'20. It comprises 60,000 parallel sentences for training, with 2,000 pairs for development, and an additional 2,000 for test.

Retrieval Settings We use the training set to retrieve analogies in both de and hsb languages. For pattern collection, we apply the Berkeley Neural Parser (Kitaev and Klein, 2018), which supports constituency parsing for de sentences. For the hsb side, we use the alignment points in hsb sentences that correspond to de patterns. To achieve this, we pre-compute sub-sentential alignments between parallel sentences in the training set, using Anymalign and Cutnalign following Lardilleux et al. (2012). Nlg is applied to gather analogies for both languages independently. Considering retrieval efficacy, we configure Nlg to produce analogical clusters with a maximum size of 5. Subsequently, we follow the instructions in Section 2.1 to construct sentence analogies.

Model and Training We use the pre-trained multilingual model `mbart-large-50`[3], which was trained to denoise corrupted texts in 50 languages, including de but not hsb. For analogy-adaptive pre-training, with the collected analogies, we take 90% of data to train the model and reserve the remaining 10% as a development set to evaluate the learning performance. The training recipe employs a batch size of 8 and incorporates early stopping: training halts if there is no improvement for three epochs. Additionally, our configurations are aligned with the settings detailed in (Liu et al., 2020), which are tailored for fine-tuning mBART models on sentence-level translation tasks.

Baselines We compare our approach with two NMT baselines, each developed using the same mBART model but differing in their continued pre-training (CP) strategies prior to fine-tuning for translation tasks. The first baseline uses the original pre-trained model, without any additional pre-training. The vanilla NMT model is directly fine-tuned on parallel sentences as a straightforward application of pre-trained language models to translation tasks. The second baseline introduces a CP phrase, where the model is updated on given sentences with the BART objective before fine-tuning. More precisely, we develop the CP baseline by further training the model with the BART pre-training scheme: masked sequence-to-sequence learning on monolingual sentences in the two languages in the training set.

3.2. Translation Results

Table 1 presents BLEU scores of translation models under three different CP strategies. For the CPoA models, we explore five different settings on coverage thresholds (λ_c) for analogies used, ranging from 0.8 to 0.4, where lower thresholds select larger but less stringent analogy sets. Compared to the vanilla NMT, models that benefit from additional exposure to the given data demonstrate improved translation quality in both directions. The baseline model, updated with CP on the entire training sentences, exhibits marginal gains of +0.4 points for de→hsb and +0.6 points for hsb→de translations. Moreover, we observe that our CPoA models show further improvements. In particular, a CPoA model with a coverage threshold of 0.8, trained on 11k

[2] `https://www.statmt.org/wmt20/unsup_and_very_low_res`

[3] `https://huggingface.co/facebook/mbart-large-50`

Model	CP data			de-hsb	
	λ_c	# (de)	# (hsb)	$\rightarrow$	$\leftarrow$
w/o CP	-	-	-	52.0	51.1
w/ CP	-	60,000	60,000	52.4	51.7
w/ CPoA	0.8	6,306	5,040	53.2	**52.4**
	0.7	78,112	48,944	**53.4**	51.1
	0.6	131,992	56,470	52.8	51.7
	0.5	200,725	63,988	52.8	51.3
	0.4	385,582	91,249	50.8	50.7

Table 1: BLEU scores of mBART models fine-tuned on a dataset of 60k parallel sentences for translations between de and hsb, with confidence intervals for these scores approximately at 1.1. For models with continued pre-training, we outline the number of training instances (sentences or analogies), along with coverage thresholds λ_c that dictate the quality of analogies selected.

Model	CP data			de-hsb	
	λ_c	# (de)	# (hsb)	$\rightarrow$	$\leftarrow$
w/o CP	-	-	-	39.0	38.6
w/ CP	-	10,000	10,000	41.3	39.4
w/ CPoA	0.8	31	1	41.7	**40.2**
	0.7	3,527	1	41.7	39.9
	0.6	4,268	4	**41.9**	40.0
	0.5	5,197	18	**41.9**	39.6
	0.4	10,954	152	41.4	39.0

Table 2: Comparison of translation performance in BLEU of mBART models in a limited scenario with only 10k training parallel sentences. The confidence intervals for BLEU scores are around 1.2.

($\approx$ 6,306+5,040) close analogies, where at least 80% of the quadruple content forms strict analogies in form, achieves an additional increase of +0.8 and +0.7 points over the model with CP. It is noteworthy that this superior performance is achieved while using less than 10% of the data used by the CP baseline. This underscores the remarkable effectiveness of the CPoA approach, which elicits knowledge from the original data using high-quality analogies, boosting translation performance with considerably less data for updates.

As the threshold decreases, expanding training data with more diverse analogies, there is potential for enhanced performance despite the inclusion of less stringent analogies. However, there is a caveat: poor analogies could lead to degradation. Specifically, when $\lambda_c = 0.4$, a large number of analogies are identified, but approximately half of these are far analogies, where less than half of the content contributes to forming a strict analogy. Such poor analogies can detrimentally impact translation performance, even falling below the performance of the vanilla NMT baseline. This probably highlights the precedence of analogy quality over mere quantity in optimizing translation.

3.3. A Limited Scenario

We also conduct experiments in a limited scenario, where there are only 10,000 parallel examples for training. As shown in Table 2, our approach consistently outperforms the original model without CP. It is striking that with the update of using just 32 analogies, there is a substantial improvement in translation quality, achieving gains of +2.7 and +1.6 BLEU points.

In addition, we observe that there are few Upper Sorbian analogies employed to train the mBART model. Our models mainly learnt from German analogies have shown improvements in translations both from and into Upper Sorbian, despite minimal exposure to Upper Sorbian analogies during continued pre-training. This reveals the potential influence between languages. Our retrieved analogies, albeit approximate, capture analogical regularities between sentence pieces on forms. To some extent, formal analogies can transcend specific linguistic boundaries, exhibiting linguistic-agnostic generalization. In other words, the inferential capabilities learned from analogies in one language can be generalized to others. Therefore, we speculate that learning to solve analogies can impact the language representations of multilingual models, which bolsters the model's performance in translations in both directions.

3.4. Cognitive Evaluation

To investigate the analogical ability, we evaluate the models updated with CPoA using 2,000 perfect analogies[4] between patterns, evenly distributed between de and hsb. By formatting each analogy question as a masked input $A : B :: C :$`[mask]`, we task the model with generating a solution to complete the analogy by filling in the masked token. We compute the Levenshtein distance between generations and references in characters. The results are presented in Table 3.

A critical finding from this investigation is the positive correlation between the amount of analogies and improved performance in solving analogies. Specifically, when the quality of analogies is held constant (by setting the same coverage threshold), we observe that models trained with a larger amount of analogies retrieved from the 60k set yield lower average distances compared to those applied on the 10k set. Moreover, the quality of analogies can significantly influence model performance. When the quantity of analogies is approximately the same, we observe that closer analogies with

[4] These analogies are sampled from intermediate results of pattern analogies during retrieval.

λ_c	# k	Error distance in characters de	hsb	avg
60k parallel data				
0.8	11	$6.1_{\pm 0.6}$	$4.1_{\pm 0.5}$	$5.1_{\pm 0.4}$
0.7	127	$10.4_{\pm 2.0}$	$8.5_{\pm 0.8}$	$9.5_{\pm 1.1}$
0.6	188	$11.4_{\pm 1.3}$	$8.6_{\pm 0.7}$	$10.0_{\pm 0.7}$
0.5	265	$13.7_{\pm 2.2}$	$10.3_{\pm 0.8}$	$12.0_{\pm 1.2}$
0.4	477	$14.4_{\pm 1.1}$	$12.8_{\pm 0.9}$	$13.6_{\pm 0.7}$
10k parallel data				
0.8	0	$42.2_{\pm 2.2}$	$53.7_{\pm 3.8}$	$48.0_{\pm 2.2}$
0.7	4	$9.8_{\pm 1.2}$	$9.8_{\pm 1.0}$	$9.8_{\pm 0.8}$
0.6	4	$12.4_{\pm 1.3}$	$17.0_{\pm 1.4}$	$14.7_{\pm 1.0}$
0.5	5	$19.9_{\pm 1.6}$	$6.9_{\pm 0.8}$	$13.4_{\pm 1.0}$
0.4	11	$34.9_{\pm 2.0}$	$18.6_{\pm 1.4}$	$26.8_{\pm 1.3}$

Table 3: Performance of CPoA models trained on various sets of analogies retrieved in two data scenarios in solving formal analogies. The number of analogies is approximated to the nearest thousand, representing the aggregate of German and Upper Sorbian analogies as detailed in Table 1 and Table 2. The average lengths of reference answers for these analogies are 53 characters for de and 31 characters for hsb, respectively. We measure the Levenshtein distance in characters, which indicates errors.

high quality are more effective in enhancing the accuracy in generating appropriate solutions. This effect is particularly evident when comparing the performance of models trained with 11k analogies, where the distinction in error measurements is directly attributed to the quality of analogies used in CPoA. However, it is possible that the model might fail to generalize from insufficient data even though it is assigned with high quality. This point is underscored by our experiments with the smaller 10k dataset; when set to a coverage threshold of 0.8, the model limited to significantly less than 1k high-quality analogies, faces difficulties in solving analogies in both languages. As the size increases to thousands of training instances, the quality of analogies emerges as a critical determinant of the effectiveness of analogical learning.

3.5. Computational Efficiency

In terms of memory consumption, CPoA uses a parameter-efficient fine-tuning strategy that minimizes memory requirements, offering an advantage over strategies that require tuning the entire model. Table 4 provides detailed insights, presenting the training time for updating the pre-trained mBART model with CPoA. In evaluations on the 60k set, especially with a coverage threshold of 0.8, CPoA demonstrates notable improvements in a remarkably time-efficient manner. This is made by leveraging high-quality analogies, which constitute only a minor portion of the entire data, allowing for enhancement within a short training time of 3.3 ks (less than one hour).

Dataset	λ_c 0.4	0.5	0.6	0.7	0.8
60k	103.1	52.7	48.1	37.9	3.3
10k	2.3	1.2	1.0	1.0	0.5

Table 4: Training time in thousand seconds (ks) for running CPoA under different coverage threshold settings across two parallel datasets.

In addition, the feasibility of running CPoA is also evidenced in more constrained scenarios, such as with our 10k training set. CPoA effectively uses thousands or even fewer analogies to update pre-trained models. This process takes less than 20 minutes, yet still results in enhanced model performance in downstream tasks.

4. Conclusion

In this work, we explore the potential of using sentence analogies to improve performance of a pre-trained mBART model for low-resource translation. With small parallel data, we propose to structure sentences in analogical relations and continue training the model in solving sentence analogies. By continued pre-training, the model becomes more proficient in learning the mappings between two languages when fine-tuning on parallel sentences. Our approach achieves up to 2.9 and 1.6 BLEU point improvement in de→hsb and hsb→de directions with 10k parallel data. We also examine the trade-off between the quantity and quality of analogies. We observe that closer analogies are more conducive to enhancing cognitive ability. Having a vast number of analogies when many are less informative, can adversely affect translation performance.

Limitations Our approach relies on syntactic parsers to collect patterns from sentences. It cannot be applied to language pairs for which parsers are not available for neither side. In addition, we only conduct experiments on the de-hsb language pair. The results in other languages might be different.

5. Acknowledgements

The research reported in this paper was supported in part by a grant for Kakenhi (kiban C) from the Japanese Society for the Promotion of Science (JSPS), n° 21K12038 "Theoretically founded algorithms for the automatic production of analogy tests in NLP".

6. References

Alexis Conneau and Guillaume Lample. 2019. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Rashel Fam and Yves Lepage. 2018. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Douglas R Hofstadter. 2001. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Philippe Langlais and François Yvon. 2008. Scaling up analogical learning. In *Coling 2008: Companion volume: Posters*, pages 51–54, Manchester, UK. Coling 2008 Organizing Committee.

Phillippe Langlais. 2013. Mapping source to target strings without alignment by analogical learning: A case study with transliteration. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 684–689, Sofia, Bulgaria. Association for Computational Linguistics.

Adrien Lardilleux, François Yvon, and Yves Lepage. 2012. Hierarchical sub-sentential alignment with anymalign. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 279–286, Trento, Italy. European Association for Machine Translation.

Yves Lepage. 1998. Solving analogies on words: an algorithm. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 728–734, Montreal, Quebec, Canada. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and human intelligence*, pages 351–354.

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.

Liyan Wang and Yves Lepage. 2022. Masked prompt learning for formal analogies beyond words. In *Proceedings of the Workshop on the Interactions between Analogical Reasoning and Machine Learning (International Joint Conference on Artificial Intelligence - European Conference on Artificial Intelligence (IJAI-ECAI 2022)), Vienna, Austria, July 23, 2022*, volume 3174 of *CEUR Workshop Proceedings*, pages 1–14.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.