Multilingual and Code-Switched Sentence Ordering

Alexandre Salle
VTEX, Porto Alegre, RS, Brazil
alex@alexsalle.com

Shervin Malmasi
Amazon.com, Inc., Seattle, WA, USA
malmasi@amazon.com

Abstract

Sentence Ordering (SO) is a linguistic task which requires re-ordering of shuffled sentences into a coherent paragraph. SO has downstream applications, but also serves as a semantic probe for computational models as this capability is essential for understanding narrative structures, causal and temporal relations within texts. Despite its importance, prior research has been limited to predictable English language structures and has not thoroughly addressed the complexities of multilingual and varied narrative contexts. To fill this gap, we introduce a novel and comprehensive Multilingual Sentence Ordering task that extends SO to diverse narratives across 12 languages, including challenging code-switched texts. We have developed MULTISO, a new benchmark dataset that represents these challenges. Our findings reveal that both specialized sentence ordering models and advanced Large Language Models like GPT-4 face significant challenges with this task.

1 Introduction

Advances in Language Models (LMs) have increased focus on general language understanding through increasingly sophisticated tasks requiring a deeper understanding of meaning in text. These advances are underpinned by improved representation learning of core linguistic units (morphemes, words, sentences) via methods like subword tokenization, masked LMs, and next sentence prediction - combined with significant increases in model size. At the sentence level, the self-supervised task of re-ordering shuffled tokens and sentences to recover the original sequence has been used, e.g., in BART (Lewis et al., 2020).

Sentence Ordering (SO)[1] is a task that extends the permutation recovery approach to the paragraph level by shuffling sentence order. Originally studied outside of computational linguistics, SO has been used in studies of understanding human cognition (Delis et al., 1983), as well as language learning assessment and testing (Alderson, 2000). Along the same lines, understanding longer texts has always been an overarching goal in NLP, and SO serves as a semantic probe for assessing model understanding of causal and temporal relations, and ability to reason over longer texts.

Numerous computational approaches to SO have been explored (Lapata, 2003; Logeswaran et al., 2018). However, there are several shortcomings. To our knowledge, all SO research has been on English. Further, most work uses sentences from paper abstracts or text describing entities, and recent work has shown that these texts have similar and highly regular structures, allowing models to learn simple shallow cues that result in shortcut learning (Basu Roy Chowdhury et al., 2021).

To address these gaps, we propose a comprehensive multilingual SO task using varied narratives spanning several domains and 12 languages, including challenging code-switched passages. Our proposed multilingual SO task is depicted in Figure 1. Experiments on MULTISO, a new benchmark dataset that we have created, show that both models trained specifically for SO, as well as state-of-art LLMs (GPT-4), struggle on this task.

In sum, our contributions include:

- Proposing a novel comprehensive Multilingual Sentence Ordering task;
- Releasing MULTISO, a new public dataset to advance SO research;[2]
- Evaluating MULTISO with LMs and LLMs to establish benchmarks.

[1] Sometimes called sentence arrangement or re-ordering.

[2] https://github.com/alexandres/mso

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 308 - 313

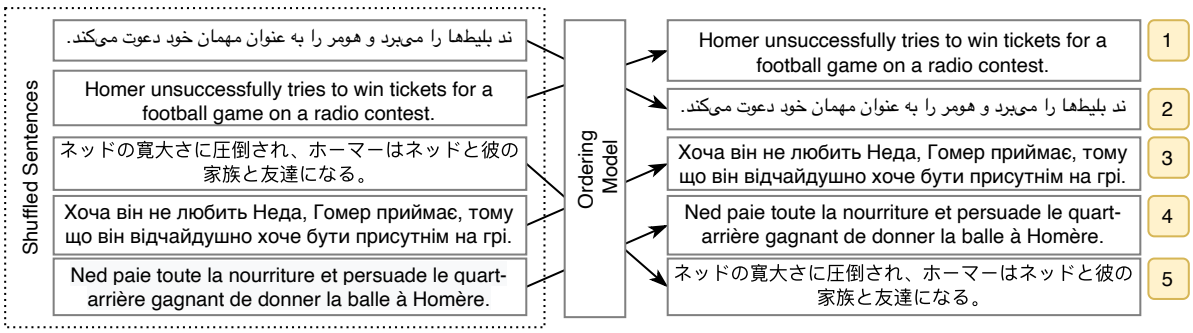

Figure 1: An example of the Code-Switched Sentence Ordering task spanning 5 languages (FA, EN, JA, UK, and FR). English versions of sentences: (2) Ned wins the tickets and invites Homer as his guest. (3) Although he dislikes Ned, Homer accepts because he desperately wants to attend the game. (4) Ned pays for all of the food and persuades the winning quarterback to give the game ball to Homer. (5) Overwhelmed by Ned's generosity, Homer becomes friends with Ned and his family.

2 Related Work

Sentence Ordering is a longstanding task within Natural Language Processing research (Lapata, 2003). SO also has more direct downstream applications in text summarization (Nallapati et al., 2017), retrieval-dependent QA (Yu et al., 2018), and concept-to-text generation (Schwartz et al., 2017). More recently, the task has gained attention with the rise of neural language models (Chen et al., 2016; Cui et al., 2018). For a comprehensive overview of the work in the area, we refer the reader to the recent survey by Shi et al. (2024).

Research on SO is nascent, and there is a paucity of benchmark tasks and datasets. Datasets such as ROC Stories (Mostafazadeh et al., 2016) provide well-structured, simple narratives composed of five sentences, purposely crafted to model coherent story progression in a strictly monolingual (English) context. Similarly, datasets based on abstracts from NIPS, ACL, and arXiv papers (Chen et al., 2016; Logeswaran et al., 2018) focus on the logical sequence of scientific ideas, yet are confined to English language scholarly texts. These datasets predominantly support tasks that require understanding simple, linear narrative structures in solely English contexts. Contrasting this, our work extends beyond the monolingual framework by introducing a novel, multilingual dataset that includes code-switching, addressing the complexities of interlaced linguistic elements. Additionally, our dataset encompasses a broader spectrum of intricate narration styles, thereby challenging models to grasp and generate more sophisticated narratives.

While SO is intrinsically interesting, it is also relevant to research on using LMs to generate semantic representations of text: Lewis et al. (2020) find that SO is an important pretraining task for downstream task performance in a monolingual setting. We hypothesize that Multilingual SO, particularly a Code-Switched variant, may help align semantic representations across languages.

Our work tries to address some of the above shortcomings by proposing a new multilingual SO task, and developing a new corresponding dataset (MULTISO) to further research in this area.

3 Multilingual Sentence Ordering

To address current gaps in the literature, we design a new SO task which is more challenging. We focus on the following areas:

- **Multilinguality**: while all previous work is on English, we expand SO to 11 new languages.
- **Challenging Data**: we work with diversely-structured narratives covering many themes.
- **Cross-lingual transfer and Code-Switching**: we define settings for zero-shot transfer, and are the first to propose mixed-language SO.

3.1 MULTISO Dataset

We have created MULTISO, a new **Multi**lingual **S**entence **O**rdering benchmark dataset[3] that includes the following monolingual, multilingual, and code-switched subtasks:

(A) **Monolingual Task**: given a shuffled narrative, the original sentence ordering must be recovered. Eight languages are included.

(B) **Cross-lingual Transfer Task**: similar to (A), but using data from 4 languages where we provide no training data (zero-shot).

[3]Available at https://github.com/alexandres/mso

(C) **Code-Switching Task**: this challenging subtask requires ordering code-switched narratives where sentences are in different languages, with up to 5 languages per story.

Examples of each task are shown in Table 1.

• The story concerns King Charlemagne, who has gotten lost and detached from his retinue in a storm. • He is forced to take refuge in the home of a collier named "Rauf". • While Rauf is more or less hospitable, he does not realize his guest is the king, and so treats him somewhat roughly.
• Родина Тернерів вирішила переселитися до штату Вірджинія. • Дорогою вони підбирають бежатнього собаку - колі Лессі. • РЛессі стає членом родини, й особливо допомагає підліткові Мету, рятуючи його в скрутних ситуаціях.
• Романтичний фільм обертався навколо рокера і глухого хлопчика. • 一人は沈黙の中で暮らし、もう一方は騒音と恐怖の中で生きている。 • The two met in a Baguio camp where hearing kids were mixed with non-hearing kids to find their common ground, which is their love for music.

Table 1: Example narratives from our data: *The Tale of Ralph the Collier* (EN), *Lassie* (UK), and *If I Knew What You Said* (Code-Switched UK+JA+EN).

Languages Our task is multilingual, spanning 12 languages: DE, EN, ES, FA, FR, IT, PT, UK, JA, SV, TR, and ZH. Detailed statistics are shown in Table 2.

Language	Train	Valid	Test	Sents/Story	Tokens/Sent
German (DE)	20k	4.6k	4.6k	5.5 ± 3.0	17.7 ± 8.2
English (EN)	20k	12.3k	12.3k	4.7 ± 2.8	20.3 ± 8.9
Spanish (ES)	20k	2.7k	2.7k	3.9 ± 2.2	22.9 ± 9.8
Farsi (FA)	5.6k	0.7k	0.7k	4.0 ± 2.5	20.9 ± 9.6
French (FR)	20k	4.7k	4.7k	3.9 ± 2.2	19.4 ± 9.1
Italian (IT)	20k	4.2k	4.2k	4.1 ± 2.4	22.2 ± 9.8
Portuguese (PT)	14.9k	1.9k	1.9k	4.0 ± 2.2	21.7 ± 9.3
Ukrainian (UK)	16.9k	2.1k	2.1k	4.8 ± 2.8	14.9 ± 7.4
Japanese (JA)	0	0.9k	7.5k	3.2 ± 1.5	55.5 ± 31.9
Swedish (SV)	0	1.3k	11.7k	4.0 ± 2.2	17.5 ± 8.1
Turkish (TR)	0	0.5k	4.7k	5.0 ± 3.0	14.5 ± 7.5
Chinese (ZH)	0	0.8k	7.2k	3.8 ± 2.1	48.0 ± 27.9
Code-Switched (CS)	20k	2.5k	2.5k	4.7 ± 2.8	16.6 ± 11.4
CS English Control (CS-EN)	20k	2.5k	2.5k	4.7 ± 2.8	20.4 ± 8.9
English Books (EN-Books)	0	0	240	6.5 ± 4.3	16.9 ± 11.1
Books Code-Switched (CSB)	0	0	240	6.5 ± 4.3	15.9 ± 10.8
Translated Books (CSB-MT)	0	0	240	6.5 ± 4.3	16.0 ± 10.7

Table 2: Per-split data statistics, with mean±std number of sentences per story (Sents/Story) and mean±std tokens (characters for JA, ZH) per sentence (Tokens/Sent). We are collecting more languages and narrative types.

Our data focuses on narratives describing stories from creative works (e.g., movies, books, TV shows). Unlike existing data used for SO (text from paper abstracts, descriptions of persons and entities), these narratives have a less regular structure, and can include any subject matter (e.g., sci-fi). Our data generation process is described below.

Monolingual Narratives (Task A) Parsing Wikipedia dumps for 12 languages, we extract narrative sections from pages of creative works. We take the first paragraph, which is often a short summary of the story with a clear start and end. We filter paragraphs that are too short (< 2 sents) or long (> 20 sents). We perform monolingual evaluation on DE, EN, ES, FA, FR, IT, PT, UK.

Cross-lingual Transfer (Task B) For JA, SV, TR, & ZH, we provide no training data and evaluate cross-lingual, zero-shot transfer.

Code-Switched Data (Task C) These are narratives where the sentences can be from up to 5 languages: EN, FR, FA, UK, & JA. As aligning the monolingual stories is noisy and challenging, we apply Machine Translation (MT) to monolingual data to create code-switched narratives.

Books Data To assess the impact of MT used in constructing Task C, we use aligned human translations of out-of-copyright books[4] in EN, DE, ES, HU, & IT to a create a Code-Switched Books (CSB) corpus. We apply MT on the English-only version of this corpus – Books (EN) – to create a MT Code-switched Books corpus (CSB-MT) for comparison to Code-Switched Books (CSB). Although this corpus is two orders of magnitude smaller than the Wikipedia-based data, its sole use here is to assess the impact of MT on the task.

Data Validation We randomly sampled 80 EN, DE, and FR monolingual narratives; 97% were found to be valid stories by native speakers.

4 Experiments and Results

Models We use the SO model from Shen and Baldwin (2021) and employ both BERT and Multilingual BERT as the underlying encoder. We also test ChatGPT models (gpt-3.5-turbo, gpt-4) in a zero-shot setting, with a prompt instructing it to order the input story. We align the output to the original story by matching generated sentences to the original input using Longest Common Subsequence (LCS), where a match between a pair of sentences (s, t) occurs when $|LCS(s,t)|/max(|s|,|t|)) \geq 0.7$. When we fail to match each sentence in the original story to a one in the generated story, we consider this a parse error and penalize the model by randomly permuting the original story to compute the metrics.

[4] https://opus.nlpl.eu/Books.php

	Monolingual (Task A)										Cross-lingual (Task B)						Code-Switched (Task C)			
	EN		DE		IT		FA		UK		EN→DE		ES→IT		EN→FA		CS		CS-EN	
	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR	τ	PMR
BERT	**0.80**	21.08	0.59	11.94	0.61	12.51	0.48	11.27	0.46	8.57	-	-	-	-	-	-	0.47	9.47	0.78	19.33
mBERT	0.78	20.26	**0.80**	21.03	**0.79**	21.03	**0.72**	20.38	**0.77**	19.80	**0.77**	**18.64**	**0.74**	**19.13**	**0.76**	**22.28**	**0.72**	16.16	**0.81**	23.04
GPT-3.5	0.39	13.88	0.30	10.45	0.37	15.61	0.25	12.60	0.25	10.53	-	-	-	-	-	-	0.16	8.96	0.39	16.30
GPT-4	0.68	**24.35**	0.68	**23.60**	0.72	**24.71**	0.66	**20.95**	0.67	**23.21**	-	-	-	-	-	-	0.58	**16.41**	0.69	**24.59**

Table 3: Pilot results for our three subtasks, using models based on BERT, Multilingual BERT, and ChatGPT (zero-shot).

	EN-Books		CSB		CSB-MT	
	τ	PMR	τ	PMR	τ	PMR
mBERT	**0.56**	**8.05**	**0.43**	**6.77**	**0.43**	**6.69**
GPT-3.5	0.06	5.64	0.04	4.50	-0.01	4.35
GPT-4	0.16	7.04	0.09	6.54	0.07	6.16

Table 4: Results on English Books, Code-switched Books (CSB), and Translated Code-Switched Books (CSB-MT).

Metrics We utilize two standard metrics from the SO literature: (1) *Kendall's Tau* (τ) (Kendall, 1938) which measures the correlation between the correct and predicted orderings in terms of inversions; and (2) *Perfect Match Ratio* (PMR) which is the proportion of predicted orderings which are absolutely correct (equal to the correct ordering). As evidenced in Table 2, sentence counts per story vary greatly with language. To control for this and allow for direct comparison between language results, rather than averaging τ and PMR across all stories, we stratify narratives by length and compute mean τ and PMR across strata, and finally compute an unweighted mean over strata means.

4.1 Main Results

Pilot results from all models on a subset of languages are shown in Table 3. We leave evaluation on all languages for future work.

Monolingual Performance (Task A) We trained BERT and mBERT models for 5 languages. mBERT has reasonable results for all languages, with higher resource languages performing better. The monolingual BERT model performs poorly on non-EN languages, demonstrating the need for multilingual (or monolingual in the target language) encoders. Overall performance on our data is much lower than existing work leveraging narrative text such as ROCStories (Mostafazadeh et al., 2016), where reported PMRs can exceed 80% (Basu Roy Chowdhury et al., 2021). This highlights the relative difficulty of our dataset.

Cross-lingual Transfer (Task B) We apply zero-shot transfer between typologically similar and diverse languages. Transfer between similar source-target pairs (EN→DE, ES→IT) achieves similar results as monolingual models: the drop in metrics is under 10%. Interestingly, training on high-resource EN data and testing on low-resource FA data increases performance over the monolingual FA model, which has a much smaller training set. This finding demonstrates that cross-lingual transfer works well for SO.

Code-switched Performance (Task C) We create a code-switched corpus (CS) where each narrative can have up to 5 languages. This data is translated from EN, and we retain the original monolingual data as a control set (CS-EN). The Code-Switched results show that it is indeed the most challenging setting, with a 30% drop in PMR and an 11% drop in τ compared to the equivalent non-code-switched corpus (CS-EN). This result is not surprising as code-mixed tasks are usually much more difficult (Fetahu et al., 2021; Malmasi et al., 2022), but it highlights that using a pretrained multilingual Transformer model is a weak baseline, and possible efforts to address this create interesting new research directions in semantic, multilingual sentence and document representation.

ChatGPT To control for costs, we sampled 500 stories from each dataset using stratified sampling by number of sentences (to better match our metrics which are macro averaged by number of sentences). Surprisingly, despite significant prompt engineering effort, GPT-3.5 struggles on all data. In contrast, GPT-4 has the highest PMR on all datasets. Interestingly, its τ is lower than both BERT and mBERT, indicating an all or nothing approach to the task: its high PMR shows that it tends to get the ordering correct more frequently than other models, but when it fails, it is a complete failure (this all-or-nothing effect is even more pronounced for GPT-3.5; in Task A IT, it has a higher PMR than BERT, but a τ nearly 40% lower). Given its difficulty, we hope further experiments with our dataset will shed some light on the degree to which SO is emergent in LLMs (Wei et al., 2022).

4.2 Impact of Translation (Books)

Table 4 shows translation does not impact SO performance; this matches our observations in validating the translated narratives. Results on Books are lower than all results in Wikipedia: this is due to different and much more varied narrative structure, domain shift, and longer sentences.

4.3 LLM Memory Test

It is reasonable to expect LLMs such as GPT-3.5/4 to be able to recall from memory plots from Wikipedia and out-of-copyright books which are used in our dataset. We test this by randomly sampling 50 EN-plot and 50 EN-books containing at least 10 sentences. We then prompt both models with the first 5 sentences of each plot in order, and check whether they are able to recall the next 5 sentences in the correct order. Generated sentences are matched to the original sentences using the same LCS technique described in section 4. Surprisingly, GPT4 is only able to recall 14/100 instances (all correctly recalled stories are from Books), and GPT3.5 even fewer, only 3/100 (also from Books).

Focusing on GPT-4, we tested whether it could perform the SO task on the 14 stories it *can* recall perfectly from memory. It fails to do so, with a PMR of 0.0 and a Kendall's Tau close to 0. Recent work shows that LLMs suffer from the Reversal Curse (Berglund et al., 2023): GPT-4 is able to answer "Who is Tom Cruise's mother? [A: Mary Lee Pfeiffer]" but fails to answer the reverse "Who is Mary Lee Pfeiffer's son?". This might be connected to the failure in the memory test: GPT4 can recall the stories if prompted in the original order, but runs into a failure when prompted out of order. Further investigation is needed to understand the cause of this failure.

5 Conclusion and Future Work

We proposed a multilingual SO task and dataset, and showed that it is challenging, particularly for code-switched data. Although based on a well explored monolingual SO task, our research is the first to address the gap in research covering non-English and code-switched languages.

Our MULTISO dataset uses narratives describing stories from creative works, making it varied and providing a challenge for language models. This dataset is the first to explore the multilingual and mixed-language directions. We expect this task and data will facilitate research in several areas. The task enables evaluation of LM representations and model reasoning over longer language units and sequences. Each task also covers multiple languages, making it possible to study cross-lingual transfer using MULTISO.

In future work, we plan to: (1) expand the dataset with more languages and narrative types to further provide researchers with valuable resources for enhancing multilingual language models (2) perform a deeper investigation on using models to solve the task, in particular LLMs.

Ethics Statement

In accordance with the ACM Code of Ethics and Professional Conduct, our work adheres to the principles of respecting privacy and honoring confidentiality by ensuring that the data used complies with the licenses of the original sources (CC-BY-SA for Wikipedia and out-of-copyright for Books) (§1.6 and 1.7). Furthermore, our study confirms that the data does not contain any personal information or harmful content (§1.2), thereby avoiding potential harm and minimizing negative consequences. We strive to maintain high standards of professional competence, conduct, and ethical practice (§2.2) throughout our research. Our commitment to ethical conduct also involves transparency and full disclosure of our data sources and limitations (§1.3). By following these ethical guidelines, we aim to contribute to the public good and uphold the principles of responsible computing (§1.1).

References

Charles J Alderson. 2000. *Assessing Reading*. Cambridge University Press.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is Everything in Order? A Simple Way to Order Sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349.

Dean C. Delis, Wendy Wapner, Howard Gardner, and James A. Moses. 1983. The Contribution of the Right Hemisphere to the Organization of Paragraphs. *Cortex*, 19(1):43–50.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1677–1681, New York, NY, USA. Association for Computing Machinery.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Aili Shen and Timothy Baldwin. 2021. A simple yet effective method for sentence ordering. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 154–160, Singapore and Online. Association for Computational Linguistics.

Yunmei Shi, Haiying Zhang, Ning Li, and Teng Yang. 2024. An overview of sentence ordering task. *International Journal of Data Science and Analytics*, pages 1–18.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.