CCAE: A Corpus of Chinese-Based Asian Englishes

Yang Liu, Melissa Xiaohui Qin, Long Wang, and Chao Huang(✉)

University of Science and Technology Beijing, Beijing, China
{lwang,chaohuang}@ustb.edu.cn

Abstract. Language models have been foundations in various scenarios of NLP applications, but it has not been well applied in language variety studies, even for the most popular language like English. This paper represents one of the few initial efforts to utilize the NLP technology in the paradigm of World Englishes, specifically in creating a multi-variety corpus for studying Asian Englishes. We present an overview of the **CCAE** — **C**orpus of **C**hinese-based **A**sian **E**nglish, a suite of corpora comprising six Chinese-based Asian English varieties. It is based on 340 million tokens in 448 thousand web documents from six regions. The ontology of data would make the corpus a helpful resource with enormous research potential for Asian Englishes (especially for Chinese Englishes for which there has not been a publicly accessible corpus yet so far) and an ideal source for variety-specific language modeling and downstream tasks, thus setting the stage for NLP-based World Englishes studies. And preliminary experiments on this corpus reveal the practical value of CCAE. Finally, we make CCAE available at https://huggingface.co/datasets/CCAE/CCAE-Corpus.

Keywords: Web Corpora · World English · Language Model · Data-centric AI

1 Introduction

Natural language process (NLP) has achieved significant advances with the deep learning approaches in the domain of language modeling (LM), specifically the second-generation pre-trained language models (PLMs) [1] such as BERT [2], T5 [3], and GPT-3 [4], which are based on transformer backbone [5]. PLMs are fine-tuned to the target languages or the tasks at hand, so other researchers do not have to perform expensive pre-training. Due to their advanced generalization performance, PLMs have been utilized in a wide range of downstream applications, such as machine translation [5], text classification [6], and question answering [7]. They have also been proved fruitful in capturing a wealth of linguistic phenomena and features on levels of morphology [8], lexis [9, 10], and syntax [11]. Meanwhile, they can also be applied in relevant tasks such as variety detection [12] and lexical variation identification [13].

World Englishes has become a robust field of inquiry as scholars pursue more nuanced understandings of linguistic localization and multilinguals' negotiations of language differences [14]. However, there have been few attempts to investigate various indigenized Englishes by means of PLMs. This study represents the initial effort

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14303, pp. 614–626, 2023.
https://doi.org/10.1007/978-3-031-44696-2_48

to fill this gap by creating the first free-access supra corpus on which PLMs could be pre-trained for the Chinese and Chinese-based Asian English (CAE) varieties. While previous corpora have been built for Inner and Outer Circle varieties such as the small structured ICE [15] and large-scale GloWbE [16], there has not been a publicly accessible corpus for the Expanding Circle English [17], Chinese English [18], Chinese influenced and Chinese based varieties such as Singapore English [19]. The corpus we are introducing is going to be an important data infrastructure for Asian Englishes study.

In this paper, we present the CCAE (**C**orpus of **C**hinese-based **A**sian **E**nglishes), a suite of corpora totaling 340 million words in 448 thousand documents from six locations where Chinese-based English varieties were spoken. By Chinese-based Englishes, we mean Englishes developed in Sinophone regions where varieties of Chinese are used as a main language of communication and thus serve as the dominating indigenous language or one of the indigenous languages contacting English in the monolingual or multilingual contact setting. That is to say, Chinese is the dominating agent in the formation of the nativities Englishes or has influenced the formation of various localized Englishes. In order to form a definitive research scope, we include six regional varieties (Fig. 1) under the umbrella term of Chinese-based Englishes: Chinese mainland English (CHE), Hong Kong English (HKE), Macao English (MCE), Taiwan English (TWE), Malaysia English (MYE) and Singapore English (SGE).

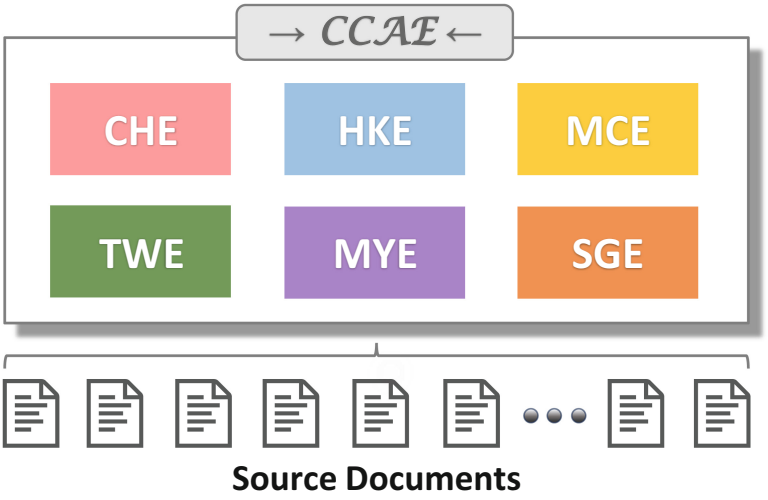

Fig. 1. Components of CCAE, totally including six varieties.

The CCAE has the following major properties:

- It is the first and largest open-access web-crawled corpus for the Chinese Englishes and Chinese-based Asian Englishes.
- It is the first and largest NLP-oriented corpus for Asian Englishes and World Englishes.

- It is a clean and deduplicated corpus in the document level. Taking into account the significance of data quality for dependent tasks, we introduce custom pipeline to conduct data cleaning.
- It maintains the traceability of each document to its origin. This level of traceability makes it possible for researchers to apply the withdrawal right of individual website owners or individual persons whose data are cited on websites and are protected by GDPR [20]. It allows researchers to systematically exclude blacklisted websites.
- It serves as the initial data resource for potential usage on downstream tasks like language variety identification, lexical variation identification, and so on.

2 Related Work

As shown in Table 1, we compare CCAE with four other corpora. Here, we simply illustrate them[1], which is web-based or manually curated.

Table 1. CCAE versus other World English corpora & WikiText. * stands for the unreported item in its bibliography, and - means "not applicable". In addition, WikiText here refers to Wikitext-103.

Corpus	GloWbE	ICE	ACE	WikiText	CCAE (ours)
Varieties (CAE)	3	2	5	0	6
Disk Size	686 MiB	400 MiB	2.1 MiB	500 MiB	2.2 GiB
Documents	134k	-	-	23.8k	448k
Tokens	142M	1.8M	420k	100M	340M
Parsing Quality	Low	High	High	High	High
Cleaning Quality	Low	High	High	High	High
Corpus Type	Web	Spoken & Written	Spoken	Web	Web
Rich Metadata	✓	✓	✗	✗	✓
Deduplicated	✗	✓	✓	✓	✓
Open Licence	✗	✗	✗	✓	✓

GloWbE. The corpus of Global Web-based English is a large-scale collection of 1.8 million web pages from 20 English-speaking countries, containing over 1.9 billion tokens. It provides linguistic annotations like PoS to support the investigation of how English is used globally.

ICE. The International Corpus of English is a collection of spoken and written English from 20 regions where English is used as the first or second language. It includes over 1,000 texts and 1,200 h of audio recordings, making it a valuable resource for studying varieties of English language use across regions and cultures around the world.

[1] Note that whatever GloWbE, ICE or ACE, they are not NLP-oriented originally, and we only counted disk size, documents and tokens on the parts of CAE in them, separately.

ACE. The Asian Corpus of English [21], an Asian English-oriented corpus capturing spoken ELF (English as a lingua franca) interactions in various regions of Asia.

WikiText-103. WikiText-103 [22] consists of over 100 million tokens extracted from English Wikipedia, it is commonly used as a benchmark dataset for training and evaluating language models. This corpus can be deemed as one of the representations of Inner-circle English.

3 CCAE at a Glance

To comprehend accurately, it is essential to understand the origin of the texts that form it. Therefore, we describe CCAE's text and metadata respectively in terms of (1) corpus-level statistics, (2) the frequency of various internet domains as text sources, and (3) the utterance date when the websites were initially indexed.

Table 2. Corpus-level statistics for CCAE.

Variety	Disk Size	Weight	Websites	Docs	Tokens	Mean Document Size
CHE	766 MiB	33.39%	145k	147.3k	114M	5.32 KiB
HKE	410 MiB	17.87%	90k	90.5k	62M	4.63 KiB
MCE	33 MiB	1.44%	9k	9.3k	5M	3.63 KiB
TWE	307 MiB	13.38%	46k	46k	42M	6.83 KiB
MYE	258 MiB	11.25%	51k	51.5k	40M	5.12 KiB
SGE	520 MiB	22.67%	103k	103.3k	77M	5.15 KiB
TOTAL	2.2 GiB		438k	448k	340M	5.24 KiB

3.1 Corpus-Level Statistics

We collected a total of 101 GB WARC(Web ARChive)[2] files for the CCAE. After document-level deduplication, the corpus is composed of 448k documents and 340M word tokens(measured by SpaCy[3] tokenization). Basic statistics of the disk size for the cleaned corpus, collected websites, documents, and tokens are displayed in Table 2.

[2] See the following Wikipedia page for more information on this standard file format:https://en.wikipedia.org/wiki/Web_ARChive.

[3] SpaCy Tokenizer: https://spacy.io/api/tokenizer.

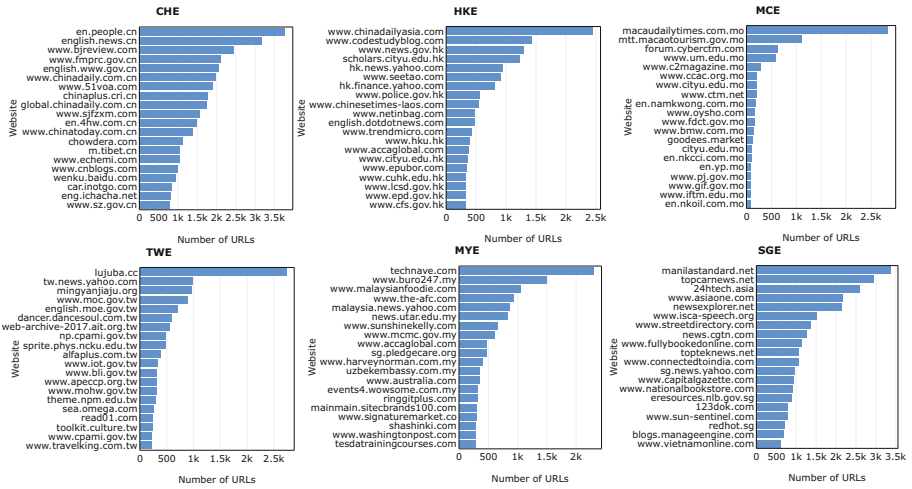

Fig. 2. Top 20 highest frequently occurring websites for each variety.

3.2 Domains Distribution

We have conducted analysis on the highest frequent top-level domains (TLD) for each variety. Predictably, most of the represented URLs are from some popular top-level domains like .com, .net, and .org. Apart from this common case, the URLs mainly consist of variety-corresponding TLD, for instance, "Chinese Mainland" has nearly 57% portion for ".cn", and "Hong Kong" has 34% portion for ".hk".

In addition, we present the top 20 highest frequently occurring websites for each variety in Fig. 2, to display the distribution of text across different websites for each variety.

3.3 Utterance Date

Language undergoes change quickly, and the accuracy of statements depends on when they were made. We attempted to determine the date of each document by examining the publish date from two sources: Google search and Internet Archive[4]. We used the earlier date as the publish date for each web page. We note that the use of the Internet Archive is not perfect, as it sometimes indexes pages months after their creation and only indexes around 65% of the URLs in CCAE. For web pages with unknown dates, we marked them as "NULL" in later storage.

As shown in Fig. 3, regardless of variety, we found that the dates of approximately 96% URLs were distributed from 2011 to 2022. In addition, there is also a significant amount of data that was written 10 years before the data collection period (from 2022/01 to 2022/06), indicating a long-tailed distribution.

[4] Internet Archive: https://archive.org/web.

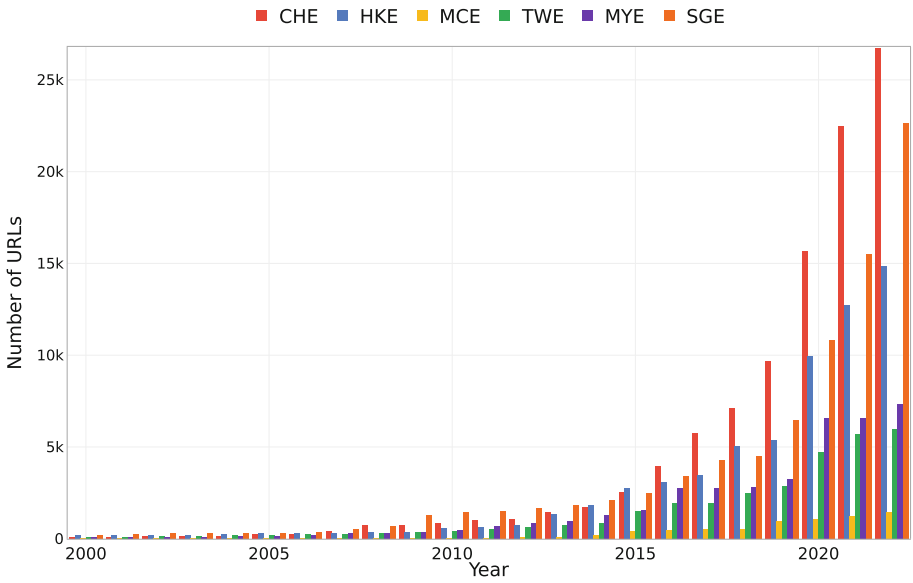

Fig. 3. The date when URLs were first indexed by the Google or Internet Archive in six Asian English Varieties

4 Generation of CCAE

4.1 Data Collection

To create a corpus available to permit research on a wide range of phenomena in six Chinese-based Englishes by performing downstream tasks in NLP (or conventional linguistic approaches), there are three principal considerations: (a) major distribution coverage, (b) variety accuracy (to ensure that the web pages were correctly associated with each of the six locations) and (c) domain diversity.

Major Distribution Coverage is crucial for creating corpora that can yield rich language features and be beneficial to models' generalization. To achieve this, we used hundreds of the highest-frequency trigrams as initial keyword seeds for query in COCA [23]. These trigrams include common three-word phrases such as "one of the" and "as well as". Algorithm 1 shows the procedure of generating a query set. We believe that using the most frequently occurring trigrams as queries is a better approach while retrieving on Google, rather than randomly selecting keywords, as it better represents the natural data distributions across different domains.

Algorithm 1 Generate trigrams as query set

Input: Corpora $C = \{d_i\}$, i = 1, ..., Sizeof(C)
Output: Trigram Query Set $\mathcal{S}$

```
L ← List()
D ← Dictionary()
for d_i ∈ C do                          ▷ iterate each document with fixed window size ≡ 3
    l, r ← 0, 3;
    L ← t ∈ Tokenize(d_i)               ▷ tokenize each document by whitespace
    while r < Sizeof(L) do
        gram ← " ".Join(L[l:r])         ▷ splice tokens in the window [l, r) with whitespace
        if gram not in D then
            D[gram] ← 0
        D[gram] ← D[gram] + 1
        l ← l + 1
        r ← r + 1
Sort(D)                                 ▷ sort D by value
S ← GetTopK(D)                          ▷ get the top-k most frequent trigrams as query set S
```

Variety Accuracy is also important to ensure that web pages are associated correctly with each of the six regions in the corpus. To achieve that, we run the trigram list we have generated against Google advanced search, in specific "search region", corresponding to each one of the varieties in China Mainland, Hong Kong, Macao, Taiwan, Malaysia, and Singapore. This method is shown to be credible [16] as Google search has adopted the following policies on web page crawling: (1) recognizing the top-level domain region (e.g., *.cn* for China Mainland, *.hk* for Hong Kong); (2) Identifying the web server's IP address; (3) Determining who links to the website; and (4) Analyzing the website's visitors, which has correctly identified website by its region. Through this method, we collected all the URLs for each item (i.e. a trigram as a query), in the result page and finally generate a deduplicated URL set for each variety.

Domain Diversity is a very important factor for the representativeness of the corpus. Collected documents should cover as many domains as possible such as technology, sport, finance, and arts. By employing human checking manually, in detail, it involves employing annotators to manually verify and validate the collected data. The annotators ensure that the collected data represents a diverse range of domains. This helps to confirm that the resulting dataset is balanced and representative of the language being studied.

We developed a collector to leverage Selenium[5] and ChromeDriver[6] to simulate human behavior, for collecting URLs. To address the obstruction of reCAPTCHA - the anti-crawled system which Google search adopts, we use 2captcha[7], a third-party online

[5] Selenium: https://www.selenium.dev.
[6] Webdriver for chrome: https://chromedriver.chromium.org.
[7] 2captcha - a captcha solving service: https://2captcha.com.

service for bypassing its verification code. To balance the requested servers' workload, we used different request proxies from around the world and keep the query frequency low to be friendly to response servers.

After the creation of the URL set for each variety, the final step of the collection is to start the downloader script to download the web page corresponding to each URL. Subsequently, we generate WARC files that contain every request and response we sent, this allows us to experiment with later processing, without hitting the server again. The crawling exclusively scraped websites whose robot files allowed it, resulting in a total of 438,625 websites across six varieties. The final raw crawling size is about 101 GB of WARC files.

4.2 Data Pre-processing

Data quality is key when building a corpus. In this section, we discuss the corpus' data pre-processing from three aspects including parsing, cleaning, and deduplication. We introduce our pipeline to accomplish the tasks of pre-processing.

In general, a web page contains different parts (e.g., header, body, tail), while we only need text body. To extract text on accuracy concern, we introduce a web text extraction tool JusText [24] which is based on a heuristic approach, to extract text from HTML pages.

After running JusText, we filter out any lines that don't end with a terminal punctuation mark. This is to guarantee that the sentences that stay are both valid and meaningful. Documents with less than five sentences are eliminated as they may not provide enough context to comprehend the text. Additionally, any unnecessary symbols or punctuations are taken out of the text, unsuitable words or words that are deemed offensive are also removed.

We noted that some of the URLs are wrongly associated with their respective regions, for example, a URL with "hk" turned out to be identified with the region of Macao (appears in the result of Macao collections). We guess it is possible as there is a quite low error rate of archive events in Google index policy, even if it follows almost credible strategies of web page categorization in the above discussion. We moved misidentified URLs to their correct class.

4.3 Output Storage Format

The output data format of CCAE we defined is JSON, there is a unique JSON document with the following data fields:

- TextID: unique document identifier consisting of an eight-digit-width integer over the whole corpus.
- Time: this field of data can be used to track changes in the variety of language use over time and to identify trends in language variation.
- Words: word count of this document.
- Variety: English variety this document belongs to.
- URL: URL address from which the web page content has been extracted.
- Title: textual title of this document.

- Content: full pre-processed text of this document.

A sample JSON document in the corpus is shown below.


```
{
    "TextID": 00019734,
    "Time": "2019-01-28",
    "Words": 741,
    "Variety": "cn",
    "Genre": "G",
    "Domain": "www.scyxxc.com",
    "URL": "http://www.scyxxc.com/en/m/news/370.html",
    "Title": "<textual title>",
    "Content": "<textual content>"
}
```


As for the text with linguistic tags, we provide another version of data to support it, each word in this replica is aligned with its multiple tags like lexeme and PoS.

5 Applications of CCAE

In this section, we demonstrate how CCAE can be used for tasks like variety-oriented language modeling and automatic variety identification, and discuss its usage for further research. We note that the utility of CCAE stretches beyond the two use cases to a wider range of language variety-relevant text mining tasks.

5.1 Multi-variety Language Modeling

Task. As one of the few trials to combine NLP with World English, we conduct a preliminary experiment on the task of language modeling. We investigate different experimental settings on multi-variety Asian English through perplexity computations by GPT-2 [25]. Through the experiments, we hope to shed light on the unique linguistic characteristics of multi-variety Asian English and the challenges it presents for language modeling.

Table 3. Zero-shot (ZS) & Fine-tuning (FT) performance for six varieties. We evaluate test perplexity (lower is better) on the validation set, for each variety, we use its abbr. to refer to itself.

CCAE	CHE	HKE	MCE	TWE	MYE	SGE	Avg
GPT-2-ZS	21.2	21.0	20.6	32.2	28.4	21.4	24.1
GPT-2-FT (mixture)	16.2	16.1	15.2	24.9	17.9	15.9	17.7
GPT-2-FT (specific)	**15.6**	**15.0**	**12.1**	**24.3**	**15.2**	**14.8**	**16.1**

Setup. The baseline model used in our experiment is the 345M GPT-2 model, implemented by Hugging Face Transformers [26] and PyTorch [27]. We utilize Adam optimization [28] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, set the dropout [29] value to 0.1, and use a learning rate of 5e-5 with batch size of 65,536 tokens. Training is conducted for a maximum of 100k iterations, with early stopping performed over the validation set. The experiments are run on 8 × NVIDIA A100-80GB GPUs.

Respectively, we conduct three settings of runs: (1) Zero-shot prompting: we simply drive the model on the validation set which we split, without any training; (2) Fine-tuning with mixed training sets across six varieties: we fine-tune the models with data of each specific English variety and test it with corresponding validation set and (3) Fine-tuning with training set of each specific variety: we merge all the parts of the training set of each variety, and training one model, and then test the model on the original split of validation set for each variety.

Results and Discussion. As shown in Table 3, We first use the original checkpoint of GPT-2 to compute the token level perplexity on each English variety directly. And then we carry out the same evaluation with the setting of supervised fine-tuning (SFT) [30]. Intuitively, the results show that the data from CCAE can significantly improve the language modeling performance on the metric of perplexity. We argue that specific SFT has considerable potential to increase the capability of "understanding" different language varieties, compared with fine-tuning with mixture of data. It indicates that the necessity of creations for variety-aware language models are nonnegligible. However, delicate experiments still need to be designed to consider the impact of variant varieties on language models with SFT in order to obtain a final credible conclusion.

5.2 Automatic Variety Identification

Task. Automatic variety identification (AVI) is a more intricate and nuanced task compared to language identification, as it demands the capability to differentiate between numerous variations of a single language, and the linguistic variations among related varieties are less apparent than those among distinct languages [31]. Consequently, it has become an appealing subject for many researchers in recent years [32,33].

Setup. CCAE naturally supports the task of AVI, thanks to its rich metadata. We present our preliminary trial on the AVI task(essentially, it is a long document classification task for our data) for six Asian English varieties in CCAE. In brief, we employ Longformer [34] and fine-tune it as an examplar baseline on our dataset, which is random-sampled in the proportion of the original distribution from CCAE. The label of each datapoint is generated by its original variety type, after carrying out a manual inspection on a randomly selected sample, we determined the precision of the label confidence, which resulted in >0.95. This validates the caliber of our supervised information and, consequently, our resources.

Results and Discussion. Not surprisingly, the results of this experiment (cf. Table 4) clearly highlight that the few-example categories seem to be more difficult to capture

Table 4. Precision, Recall and F1 for variety identification experiment on validation set.

Variety	Precision	Recall	F1	#
CHE	80.46	80.02	80.24	1,472
HKE	62.71	65.96	64.29	905
MCE	77.92	63.82	70.17	94
TWE	70.53	66.08	68.23	460
MYE	70.86	69.76	70.31	516
SGE	74.33	75.41	74.86	1,033
macro avg	72.80	70.18	71.35	4,480
weighted avg	73.28	73.16	73.19	4,480

the unique characteristics of its class. To advance significantly, it suggests that compiling larger datasets and collecting more examples for minority classes (especially for MCE), or employing more advanced models to solve this problem are considerable optimizatized directions.

6 Conclusion and Future Work

We develop CCAE, a novel multi-variety web corpora for enhancing Asian Englishes study. CCAE contains six English varieties in Asian Chinese-speaking areas. The corpus provides affluent metadata and annotations for usages. We host two versions of the data for download, to spur further research on building language variety-adaptive language models on our corpus.

In future work, we suggest an in-depth investigation of variety-specific downstream tasks like multi-variety language modeling and automatic variety identification. Conduct experiments using CCAE to analyze the effectiveness of domain adaptation between various varieties. Through our work, we hope to encourage the community to further study World Englishes, to boost non-biased and culture-diversified language modeling development.

Acknowledgement. Many thanks to Mark Davis, for his useful suggestions on data collection. We also thank the Internet Archive for providing service on the website time archive. This work was supported in part by the National Natural Science Foundation of China under Grant 62002016 and in part by the Fundamental Research Funds for the Central Universities under Grant 06500103.

Data Availability. CCAE has been released under the CC-BY-NC-ND 4.0[8](https://creativecommons.org/licenses/by-nc-nd/4.0) license on Hugging Face's website: https://huggingface.co/datasets/CCAE/CCAE-Corpus, enabling reusers to copy and distribute the material in its unadapted form, only for noncommercial purposes, while giving attribution to the creator.

References

1. Qiu, X.P., Sun, T.X., Xu, Y.G., Shao, Y.F., Dai, N., Huang, X.J.: Pre-trained models for natural language processing: a survey. SCIENCE CHINA Technol. Sci. **63**(10), 1872–1897 (2020). https://doi.org/10.1007/s11431-020-1647-3
2. Devlin, J., Chang, M., Lee, K.,Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. abs/1810.04805 (2019)
3. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 5485–5551 (2020)
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inform. Process. Systems. **33**, 1877–1901 (2020)
5. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Systems. **30** (2017)
6. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification: a comprehensive review. ACM Comput. Surv. (CSUR). **54**, 1–40 (2021)
7. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short Papers), pp. 784–789 (2018). https://aclanthology.org/P18-2124
8. Edmiston, D.: A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages. ArXiv. abs/2004.03032 (2020)
9. Espinosa Anke, L., Codina-Filba, J., Wanner, L.: Evaluating language models for the retrieval and categorization of lexical collocations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1406–1417 (2021). https://aclanthology.org/2021.eacl-main.120
10. Zhou, W., Ge, T., Xu, K., Wei, F., Zhou, M.: BERT-based Lexical Substitution. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3368–3373 (2019). https://aclanthology.org/P19-1328
11. Tran, K., Bisazza, A.: Zero-shot dependency parsing with pre-trained multilingual sentence representations. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pp. 281–288 (2019). https://aclanthology.org/D19-6132
12. Zaharia, G., Avram, A., Cercel, D., Rebedea, T.: Exploring the power of Rsomanian BERT for dialect identification. In: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties And Dialects, pp. 232–241 (2020). https://aclanthology.org/2020.vardial-1.22
13. Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., Walde, S.: Explaining and improving BERT performance on lexical semantic change detection. In: Proceedings of the 16th Conference Of The European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 192–202 (2021). https://aclanthology.org/2021.eacl-srw.25
14. Nuske, K.: "I Mean I'm Kind of Discriminating My Own People:" A Chinese TESOL Graduate Student's Shifting Perceptions of China English. TESOL Q. **52**, 360–390 (2018)
15. Kirk, J., Nelson, G.: The international corpus of english project: a progress report. World Englishes. **37**, 697–716 (2018)
16. Davies, M., Fuchs, R.: Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). English World-Wide. **36**, 1–28 (2015)
17. Berns, M.: Expanding on the expanding circle: where do we go from here? World Englishes. **24**, 85–93 (2005)
18. Xu, Z.: Chinese English: A future power?. The Routledge Handbook Of World Englishes, pp. 265–280 (2020)

19. Leimgruber, J.: Singapore English. language and linguistics. Compass. **5**, 47–62 (2011)
20. Voigt, P., Bussche, A.: The EU general data protection regulation (GDPR). A Practical Guide, 1st Ed., Cham: Springer International Publishing. **10**(3152676), 10–5555 (2017)
21. Kirkpatrick, A.: The Asian corpus of English: motivation and aims. Perspective **28**, 256–269 (2013)
22. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. ArXiv Preprint ArXiv:1609.07843 (2016)
23. Davies, M.: The corpus of contemporary American English as the first reliable monitor corpus of English. Literary Lingu. Comput. **25**, 447–464 (2010)
24. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Masarykova Univerzita, Fakulta Informatiky, Disertacnı Práce (2011)
25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI Blog. **1**, 9 (2019)
26. Wolf, T., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
27. Paszke, A., et al: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inform. Process. Syst. **32** (2019)
28. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. ArXiv Preprint ArXiv:1412.6980 (2014)
29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
30. Ouyang, L., et al.: Training language models to follow instructions with human feedback. Adv. Neural Inform. Process. Syst. **35**, 27730–27744 (2022)
31. Yang, L., Xiang, Y.: Naive Bayes and BiLSTM ensemble for discriminating between mainland and Taiwan Variation of Mandarin Chinese. In: Proceedings of the Sixth Workshop on NLP For Similar Languages, Varieties and Dialects, pp. 120–127 (2019). https://aclanthology.org/W19-1412
32. Popa, C., NullStefănescu, V.: Applying multilingual and monolingual transformer-based models for dialect identification. In: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 193–201 (2020). https://aclanthology.org/2020.vardial-1.18
33. Ceolin, A.: Comparing the performance of CNNs and shallow models for language identification. In: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 102–112 (2021). https://aclanthology.org/2021.vardial-1.12
34. Beltagy, I., Peters, M., Cohan, A.: Longformer: The long-document transformer. ArXiv Preprint ArXiv:2004.05150 (2020)