ERAS: Evaluating the Robustness of Chinese NLP Models to Morphological Garden Path Errors

Anonymous ACL submission

Abstract

In languages without orthographic word boundaries, NLP models perform *word segmentation*, either as an explicit preprocessing step or as an implicit step in an end-to-end computation. This paper shows that Chinese NLP models are vulnerable to *morphological garden path errors*—errors caused by a failure to resolve local word segmentation ambiguities using sentence-level morphosyntactic context. We propose a benchmark, *ERAS*, that tests a model's vulnerability to morphological garden path errors by comparing its behavior on sentences with and without local segmentation ambiguities. Using ERAS, we show that word segmentation models make garden path errors on locally ambiguous sentences, but do not make equivalent errors on unambiguous sentences. We further show that sentiment analysis models with character-level tokenization make *implicit* garden path errors, even without an explicit word segmentation step in the pipeline. Our results indicate that models' segmentation of Chinese text often fails to account for morphosyntactic context.

1 Introduction

When working with languages that do not mark word boundaries in their writing systems, the task of *word segmentation*, where texts are parsed into word-level tokens, is non-trivial. Nonetheless, in high-resource languages like Chinese, the state of the art in word segmentation is strong, with F1 scores close to 100% for common benchmarks (Emerson, 2005; Huang et al., 2020; Ke et al., 2020; Lin et al., 2023, *inter alia*).

Word segmentation is subject to structural ambiguity: there may be more than one way to segment a text into words or morphemes. In English, for example, the word *unlockable* can be understood to mean 'unable to be locked' (*un* + *lockable*), or 'able to be unlocked' (*unlock* + *able*). In some cases, a sequence of words or morphemes that appears ambiguous may be disambiguated by morphosyntactic context. Figure 1 illustrates this *morphological garden path* phenomenon with an example from Mandarin Chinese. In isolation, the trigram 留心机 could be segmented as either 留 'with' + 心机 'caution' or 留心 'take note of' + 机 'the machine'. In context, however, the rest of the sentence cannot be parsed grammatically if the latter segmentation is used. We therefore say that this sentence is *locally ambiguous*—the sentence as a whole is structurally unambiguous, but it contains a substring that is ambiguous in isolation.

Grammatically Valid Segmentation

学生	留	心机	处理	友人
the student	with	caution	treats	his friends

'The student treats his friends with caution.'

Ungrammatical Segmentation

学生	留心	机	处理	友人
the student	takes note of	the machine	dealing with	his friends

Figure 1: Morphological garden paths involve local ambiguity at the level of word segmentation. In the sentence above, the bigram 留心 constitutes a valid word, but segmenting it as a word renders the sentence unparsable. A valid parse is obtained if the character 心 forms a word with the following character, 机.

This paper investigates whether Chinese NLP models correctly resolve local structural ambiguities caused by morphological garden paths. If they do not, then we expect models to make *garden path errors* by parsing locally ambiguous substrings incorrectly. To that end, we propose the Evaluation of Robustness to Locally Ambiguous Segmentation (ERAS) benchmark, a synthetic test set consisting of locally ambiguous *test sentences* paired with unambiguous but otherwise identical *control sentences*. As illustrated in Figure 2, each test sentence contains a *canary word* that can only appear in a

segmentation of the sentence if a locally ambiguous substring is parsed incorrectly.

ERAS can be used in two ways. First, ERAS can evaluate word segmentation models by comparing the models' segmentations for test and control sentences. If a canary word is parsed as a word, but the equivalent segmentation error is not made for the control sentence, then we assume that a garden path error has been committed. ERAS can also evaluate sentiment analysis models that use a byte- or character-level tokenization scheme. Using a lexicon containing human-annotated word-level sentiment ratings (Wang and Ku, 2016), we ensure that canary words carry a sentiment polarity that is incongruent with the rest of the sentence. If the model's predicted sentiment for the test sentence is influenced by the sentiment of the canary word, then we assume that an *implicit* garden path error has been committed.

This paper conducts both types of evaluations using ERAS. We first show that Transformer-based (Vaswani et al., 2017; Devlin et al., 2019) and non-neural word segmentation models make garden path errors, systemically parsing canary words as words (Section 3). We show that these errors are mostly caused by a preference for resolving garden paths in a greedy left-to-right manner. We then show that Transformer-based sentiment analysis models with a character-level tokenization scheme make implicit garden path errors on ERAS (Section 4). These errors can be mitigated by introducing word boundary information to the model during pre-training or fine-tuning. Our findings point to morphological garden paths as a source of error for Chinese NLP models, especially for models that do not receive any explicit word boundary information during pre-training or fine-tuning.

Contributions. (1) We provide a benchmark, ERAS, for determining whether a Chinese NLP model explicitly or implicitly resolves local ambiguities due to morphological garden paths. (2) Using ERAS, we show that word segmentation models are vulnerable to explicit garden path errors, and sentiment analysis models are vulnerable to implicit garden path errors.

2 ERAS: A Benchmark for Detecting Garden Path Errors

ERAS consists of 203,944 pairs of test and control sentences, available in both simplified and traditional characters, synthetically generated from templates. ERAS is organized into 39 *minimal pair paradigms* (Warstadt et al., 2020), examples of which are shown in Figure 3. Each paradigm is defined by two templates, one for test sentences and one for control sentences. Within each paradigm, the two templates are identical, except that the test template contains a morphological garden path and the control template does not.

We evaluate models on ERAS by testing the null hypothesis that models perform similarly on test and control sentences. If a model performs better on control sentences than on test sentences, then we conclude that this model is susceptible to garden path errors. ERAS is designed to test two kinds of models: Chinese word segmentation models, which may incorrectly segment morphological garden paths, and sentiment analysis models, whose outputs may be influenced by the sentiment of words that could only exist if morphological garden paths are implicitly segmented incorrectly.

Test Sentence (Locally Ambiguous)

学生	留	心	机	处理	友人
the student	with	caution		treats	his friends

'The student treats his friends with caution.'

Control Sentence (Unambiguous)

学生	留	计谋	处理	友人
the student	with	shrewdness	treats	his friends

'The student treats his friends shrewdly.'

Figure 2: ERAS consists of locally ambiguous *test sentences* paired with unambiguous *control sentences*. Test and control sentences differ in terms of a three-character test site, where test sentences contain a two-character canary word whose existence renders the sentence unparsable.

2.1 Dataset Structure

As shown in Figures 2 and 3, test and control templates differ in terms of a three-character substring called the *test site*. In test templates, the first two and last two characters of the test site form words from the ANTUSD sentiment lexicon, which overlap at the middle character. One of the two words is designated as the *true word*, while the other is designated as the *canary word*. In each test template, the true word and canary word have differing sentiment labels (+, 0, or −), and only the true word may appear in a valid segmentation of the sentence. For each test template, a control template is formed by manually generating a locally unambiguous paraphrase of the test template's test site.

Left-Branching, +/– (Canary Word: 心机 'caution')

Test:	person	留心 takes note of	机动的 the motorized	noun
Control:	person	留意 takes note of	机动的 the motorized	noun

Right-Branching, –/+ (Canary Word: 留心 'to take note')

Test:	person	留 with	心机 caution	处理 treats	noun
Control:	person	留 with	计谋 shrewdness	处理 treats	noun

Figure 3: Examples of paradigms in ERAS. Each paradigm consists of two templates, one for test sentences and one for control sentences. Each paradigm also has a branching structure (left or right) and a true and canary sentiment value (+/–, +/0, –/0, or –/+). Templates contain *slots* belonging to seven possible types: concept, entity, modifier, noun, object, person, or verb. The templates shown here contain one person slot and one noun slot.

Paradigms are parameterized by the following properties of the test template.

Branching Structure. Let $x_1x_2x_3$ denote the test site of a paradigm's test template. We say that this paradigm is *left-branching* if x_1x_2 is the true word and x_2x_3 is the canary word. Otherwise, we say that this paradigm is *right-branching*.

Sentiment. The *true sentiment* of a paradigm is the sentiment label of its true word, while the *canary sentiment* is the sentiment label of its canary word. We represent the true sentiment t and canary sentiment c of a paradigm using the notation t/c.

2.2 Explicit vs. Implicit Segmentation Errors

ERAS measures susceptibility to two kinds of segmentation errors: *explicit* and *implicit*. An explicit segmentation error occurs when a model outputs word segmentation information that parses the canary word as a word. An implicit segmentation error occurs when a model does not output word segmentation information, but nonetheless behaves as though the canary word were parsed as a word. We detect implicit segmentation errors in Section 4, where we evaluate sentiment analysis models on ERAS by testing the null hypothesis that sentiment predictions for ERAS sentences are not influenced by their paradigms' canary sentiment labels.

2.3 Dataset Construction

ERAS was constructed via the following steps.

Creation of Templates. We form an initial list of test sentence test sites by extracting all word pairs of the form (x_1x_2, x_2x_3) from the ANTUSD lexicon, where x_1, x_2, x_3 are single characters, such that x_1x_2 and x_2x_3 have different sentiment labels. For each word pair, we manually generate two Mandarin Chinese sentences subject to the following criteria: (1) both sentences contain the test site $x_1x_2x_3$, (2) both sentences are locally ambiguous at the test site and nowhere else, and (3) one sentence is left-branching and the other is right-branching. We discard all the word pairs for which we were unable to generate the two sentences, and we discard sentences for which the true sentiment label is 0.[1] The remaining 39 sentences are manually converted into test templates by replacing certain content words with one of the following *slots*: concept, entity, modifier, noun, object, person, or verb. For each test template, we construct the corresponding control template by paraphrasing the test site without local ambiguity.

Sentiment Labeling. ANTUSD provides sentiment scores on a continuous scale from −1 (negative) to 1 (positive), as well as categorical labels (positive, negative, or neutral) from each annotator. We consider a word to be labeled as + (resp. –) if its sentiment score is at least .4 (resp. less than –.4), and more annotators labeled the word as positive than negative (and *vice versa*). We consider a word to be labeled as 0 if it meets the following criteria: its sentiment score is 0, the majority of annotators labeled it as *neutral*, and at least two annotators marked it as a "non-sentiment word."

Slot Filling. Each of the seven slot types (concept, entity, modifier, noun, object, person, verb), is associated with a manually constructed word list. To generate test and control sentences, we present each template to `bert-base-chinese`[2] (Devlin et al., 2019), with all slots masked out. We then fill the slots iteratively from left to right, at each step using all words from the appropriate word list whose masked language modeling probability

[1] As explained in Subsection 2.4, we do not allow true sentiment labels to be 0 because such paradigms are not compatible with our human evaluation methodology.

[2] https://huggingface.co/google-bert/bert-base-chinese

scores exceed a certain threshold.[3]

Conversion to Traditional Characters. ERAS is constructed using simplified characters. We create a traditional-character version of ERAS using the `chinese-converter` Python package.[4]

2.4 Human Evaluation

Our setup for detecting implicit segmentation errors using ERAS is based on the assumption that both test and control sentences have the sentiment value given by the paradigm's true sentiment, but that test sentences are treated as having the canary sentiment if and only if they are segmented incorrectly due to garden path errors. In order to verify this, a sample of 20 sentence pairs from each paradigm was evaluated by 5 native speakers of Mandarin Chinese, with each annotator evaluating 4 sentence pairs per paradigm. We assume that (1) the sentiment of control sentences is given by their paradigms' true sentiment labels and (2) annotators do not make garden path errors while reading test sentences,[5] and we seek to verify that the sentiment of each test sentence matches that of its corresponding control sentence.

Annotation Task. Annotators were asked to rank sentence triples in order of sentiment polarity. For each test–control pair (t, c), annotators were presented with triples of the form $\{t, c_1, c_2\}$ and $\{c, c_1, c_2\}$, where c_1 and c_2 are control sentences. Each annotator ranked 208 sentence triples, the minimum necessary to ensure that all test–control pairs assigned to the annotator are represented.

Evaluation. We assume that any sentence that is ranked more positively (resp. negatively) than a control sentence with a label of + (resp. −) also has a label of + (resp. −). Therefore, we say that a test–control pair (t, c) with a true sentiment of + (resp. −) is *annotated correctly* if t is ranked at least as positively (resp. negatively) as c, relative to c_1 and c_2. Because we have ensured in Subsection 2.3 that true sentiment labels can only be + or −, we do not need to consider cases where the control sentence has a sentiment of 0; in such cases, unless t has exactly the same ranking as c, we cannot determine whether or not t has a sentiment of 0.

[3]The threshold is .74 for the first slot in each template, and for each subsequent slot, the threshold is increased by .05.

[4]`https://pypi.org/project/chinese-converter/`

[5]In order to ensure this, we give annotators unlimited time to complete the task, and do not require them to label examples in order.

Result. We calculate a *human accuracy* of 91.3, defined as the percentage of test–control pairs that have been annotated correctly. For 81.5% of pairs, test and control sentences received the exact same sentiment ranking. This confirms that test and control sentences have the same sentiment value in the absence of garden path errors. We use this result to calculate a human baseline in Section 4.

3 Explicit Garden Path Errors

Our first experiment evaluates word segmentation models according to their vulnerability to explicit segmentation errors due to garden path phenomena. Under the hypothesis that word segmentation models may segment canary words as words, we predict that test-sentence test sites will be segmented less accurately than control-sentence test sites.

3.1 Experimental Setup

We test word segmentation models by using them to segment all sentences in the ERAS dataset. We evaluate models based on whether test sites are segmented correctly; we do not measure how well models segment other parts of the sentence. We consider a left-branching sentence with test site $x_1x_2x_3$ to be *segmented incorrectly* if a word boundary is placed between x_1 and x_2, but not between x_2 and x_3. A right-branching sentence is segmented incorrectly if a word boundary is placed between x_2 and x_3, but not between x_1 and x_2.

Evaluation. For each paradigm, we define a model's *paradigm test accuracy* to be the percentage of test sentences within that paradigm that the model segments correctly, and we define the model's *overall test accuracy* to be the mean of the model's paradigm test accuracy across all paradigms. *Paradigm control accuracy* and *overall control accuracy* are defined analogously.

Models. We fine-tune `bert-base-chinese` models on the following word segmentation benchmarks: AS, CityU, PKU, and MSR (Emerson, 2005). Each of our models performs close to state of the art results reported by Huang et al. (2020) and Tian et al. (2020).[6] We evaluate the AS and CityU models on the traditional-character version of ERAS, and PKU and MSR on the simplified-character version. In addition to our

[6]Our AS, CityU, PKU, and MSR models obtain F1 scores of 96.7, 97.8, 96.5, and 98.3 on their respective test sets, compared to state of the art F1 scores of 96.6, 97.9, 96.6, and 98.4, respectively.

	Overall			Left-Branching			Right-Branching		
	Test	Control	Diff.	Test	Control	Diff.	Test	Control	Diff.
Transformer Models									
AS (hant)	82.3	97.4	15.1	91.4	**100.0**	8.6	76.6	89.6	13.0
CityU (yue)	81.8	**99.3**	17.5	83.2	**100.0**	16.8	**92.0**	**99.8**	7.8
PKU	84.4	94.9	10.5	94.2	**100.0**	5.8	82.4	84.6	2.2
MSR	**88.6**	94.4	**5.8**	**95.8**	**100.0**	**4.2**	88.3	89.6	**1.3**
Non-Neural Models									
Jieba	68.9	**92.3**	23.4	71.2	92.6	21.4	65.8	**91.9**	26.1
PKUSEG	**75.4**	90.0	**14.6**	**78.4**	**95.5**	**17.1**	**71.2**	82.1	**10.9**
Baseline									
MaxMatch	64.1	100.0	35.9	82.6	100.0	17.4	37.5	100.0	62.5

Table 1: Overall test and control accuracies on ERAS for word segmentation models trained on traditional Mandarin (hant), traditional Cantonese (yue), or simplified Mandarin corpora (all others). Results are reported for the entire ERAS dataset ("Overall"), along with results for left-branching and right-branching paradigms only. The MSR model is least susceptible to garden path errors. All models except for CityU perform better on left-branching test sentences than right-branching test sentences.

fine-tuned models, we also evaluate the Jieba[7] and PKUSEG (Luo et al., 2022)[8] non-neural word segmentation models.

Baseline. Our baseline for this experiment is the MaxMatch algorithm (Chang et al., 2008), which iterates through a text from left to right and greedily segments the longest string starting at the current position that exactly matches at least one word from the MSR training corpus.

3.2 Results

As reported in the first column of Table 1, all of our models achieve a lower accuracy on test sentences than control sentences. This confirms our hypothesis that garden paths serve as a source of explicit word segmentation errors. We find that the MSR and PKU models are the least susceptible to garden path errors, in the sense that they exhibit the smallest gap in performance between test and control sentences. By that metric, the MaxMatch baseline and the Jieba model are the most susceptible to garden path errors. The non-neural PKUSEG model slightly outperforms the AS and CityU models in terms of test–control gap, but this is largely because PKUSEG achieves the lowest overall accuracy on control sentences.

Left-Branching Bias. The other columns of Table 1 report overall accuracies for left- and right-branching paradigms, separately. These results reveal a bias in our models for producing left-branching segmentations, in the sense that all of our models perform better on left-branching sentences than right-branching sentences. The only exception to this trend is that CityU performs better on right-branching test sentences than left-branching test-sentences. MaxMatch explicitly incorporates this bias by iterating through the text from left to right; but since the other models do not incorporate an explicit concept of directionality, we assume that the left-branching bias in these models is learned from the training corpus.

4 Implicit Garden Path Errors

Our second experiment evaluates sentiment analysis models according to their vulnerability to implicit segmentation errors due to garden path effects. Our hypothesis is that the sentiment polarity of canary words will contribute to the sentiment labels assigned by the models to test sentences, even though these words should, in theory, be excluded from the semantics of these sentences. We therefore predict that, for example, a sentiment analysis model should label a test sentence with a negative-sentiment canary word as being "more negative" on average than the corresponding control sentence. We also predict that providing word boundary information during pre-training or fine-tuning will improve a sentiment model's performance according to the aforementioned metric.

4.1 Experimental Setup

We test binary sentiment analysis models, which produce a single logit representing predicted sentiment polarity, by having them predict the sentiment of each sentence in ERAS. Our evaluation of these

[7] https://github.com/fxsjy/jieba
[8] https://github.com/lancopku/pkuseg-python

Occluded Test Sentence

学生	[MASK]	心 机	处理	友人
the student		caution	treats	his friends

'The student treats his friends with caution.'

Occluded Control Sentence

学生	[MASK]	计谋	处理	友人
the student		shrewdness	treats	his friends

'The student treats his friends shrewdly.'

Figure 4: Our occlusion study ablates canary words by masking out ("occluding") the character in the test site not belonging to the true word.

models is analogous to the paradigm described in Subsection 2.4, which we used for our human evaluation of ERAS's sentiment properties. To that end, we consider a +/– or +/0 sentence pair to be *classified incorrectly* if its test sentence receives a lower sentiment score than its control sentence. A –/+ or –/0 sentence pair is classified incorrectly if its test sentence receives a higher sentiment score than its control sentence.

Evaluation. We define the *paradigm accuracy* of a sentiment analysis model on a paradigm to be the percentage of sentence pairs in that paradigm that are classified correctly by the model. The *overall accuracy* of a model on ERAS is the mean of that model's paradigm accuracies across all paradigms.

Causal Analysis. We expect that a model that is completely immune to implicit garden path errors would make no systematic distinctions in sentiment between test and control sentences. Such a model would achieve an overall accuracy of 50%. When a model achieves an overall accuracy of 50%, however, we do not know whether this is because the model did not make any garden path errors, or whether it is because the model made some number of garden path errors while also making errors in the opposite direction. For this reason, we perform an *occlusion study* (Zeiler and Fergus, 2014) in order to determine what percentage of ERAS misclassifications were caused by garden path errors.

As illustrated in Figure 4, our occlusion study involves masking out the 1st character in right-branching test sites and the 3rd character in left-branching test sites, effectively ablating the canary word in test sentences. We assume that an incorrect ERAS classification is caused by an implicit garden path error if ablating the canary word causes the model's control and test outputs to become more similar to one another. Following Balkir et al. (2022), we define *necessity* to be the percentage of ERAS errors that are caused by implicit garden path errors according to our occlusion test, and *sufficiency* to be the percentage of implicit garden path errors detected by our occlusion test that result in an ERAS misclassification. We calculate the *garden path error rate* (GPER) as follows:

$$\text{GPER} = (1 - \text{overall accuracy}) \cdot \text{necessity}.$$

The GPER measures the percentage of ERAS examples for which the model makes an implicit garden path error, leading to a misclassification.

Models. We evaluate three off-the-shelf sentiment analysis models from the Erlangshen family (Zhang et al., 2023): RoBERTa-110M[9], RoBERTa-330M,[10] and Megatron-BERT-1.3B.[11]

To test our prediction regarding word boundary supervision, we consider two methods for injecting word boundary information into a sentiment analysis model: *whole word masking* (Cui et al., 2021), where only full words are masked during masked language model pre-training, and joint fine-tuning on word segmentation. We fine-tune sentiment analysis models with whole word masking ("WWM"), joint training on word segmentation ("CWS"), and a combination of these two techniques ("WWM+CWS"), and compare these models to a model fine-tuned on sentiment analysis only without whole word masking ("Sentiment Only"). Our WWM and WWM+CWS models are fine-tuned from the `chinese-bert-wwm-ext` model,[12] which was pre-trained with whole word masking, while our CWS and Sentiment Only models are fine-tuned from `bert-base-chinese`. Our CWS and WWM+CWS models are jointly fine-tuned on the MSR dataset.

All of our models use character-level tokenization, and none of them incorporate an explicit word segmentation preprocessing step, apart from whole word masking or joint tuning on word segmentation.

[9] https://huggingface.co/IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment

[10] https://huggingface.co/IDEA-CCNL/Erlangshen-Roberta-330M-Sentiment

[11] https://huggingface.co/IDEA-CCNL/Erlangshen-MegatronBert-1.3B-Sentiment

[12] https://huggingface.co/hfl/chinese-bert-wwm-ext

	Overall		Occlusion		Control – Test Sentiment				Test Set	
	GPER	Acc.	Necc.	Suff.	+/–	+/0	–/0	–/+	ASAP	ChnSentiCorp
Off-the-Shelf Models										
RoBERTa-110M	45.4	41.7	77.9	89.4	6.3	4.9	**2.3**	–18.7	97.8	96.6
RoBERTa-330M	**38.0**	**52.2**	79.6	**50.3**	**1.4**	**.9**	6.0	**–2.2**	97.9	96.7
Megatron-1.3B	41.8	43.8	**74.3**	68.7	2.1	13.6	8.3	–15.9	**98.1**	**97.0**
Fine-Tuned Models										
Sentiment Only	36.0	**49.5**	71.3	74.9	**–3.3**	**–1.9**	**.3**	–27.6	**97.6**	95.4
WWM	32.6	50.8	66.2	67.7	5.5	–5.0	–6.8	–11.0	97.5	95.1
CWS	23.7	41.3	**40.3**	45.9	–9.7	9.2	–10.9	–20.3	97.2	**95.5**
WWM+CWS	**18.8**	59.8	46.8	**34.5**	–5.9	–6.5	**.3**	**.6**	97.0	**95.5**
Baseline										
Humans	8.7	50.6								

Table 2: Garden path error rates (GPER, smaller is better) and overall accuracies (closer to 50 is better) attained by our sentiment analysis models, compared to a human baseline. Test set accuracy (higher is better) and occlusion study results (lower is better) are also reported. "Control – Test Sentiment" reports the average difference in probability scores for the + class between control and test sentences (closer to 0 is better), where positive values for +/– and +/0 and negative values for –/+ and –/0 constitute evidence of implicit garden path errors. Among the off-the-shelf models, RoBERTa-330M is least susceptible to garden path errors. GPER results indicate that a combination of whole word masking and joint training on word segmentation (WWM+CWS) is effective in reducing susceptibility to garden path errors, though the accuracy and test set results suggest that the latter makes the model more susceptible to errors caused by other factors.

Baseline. We compare our models against the results of the human evaluation described in Subsection 2.4. The human accuracy score of 91.3 reported there is not directly comparable to accuracies for sentiment analysis models, however, because the sentiment values generated by the annotators are discrete rather than continuous. In order to make the human evaluation results comparable to the results of this experiment, we assume that half of the 81.5% of ERAS sentence pairs where test and control sentences received the same sentiment ranking would be annotated correctly in the continuous setting, and that all misclassifications are caused by garden path errors. This leads to a human baseline accuracy of 50.6 and GPER of 8.7.

4.2 Results

The results of our implicit garden path error experiment are shown in Table 2. The RoBERTa-330M model achieves the best GPER among the off-the-shelf models, and our WWM+CWS model achieves the best GPER among our fine-tuned models. The RoBERTa-330M model also achieves the best overall accuracy among the off-the-shelf models, but the WWM+CWS model achieves the worst overall accuracy across all models. Instead, the Sentiment Only model achieves the best overall accuracy, outperforming the human baseline.

Error Analysis. The right-hand side of Table 2 ("Control – Test Sentiment") shows the average difference in sentiment scores assigned by our models to control vs. test sentences, where scores range from 0 (most likely to be –) to 100 (most likely to be +). We would ideally like these numbers to be as close to 0 as possible, since this would indicate that the model makes no systematic distinction between control and test sentences. This error analysis reveals that for all models except RoBERTa-330M and WWM+CWS, the –/+ paradigms result in the greatest discrepancy in output between control and test sentences. Indeed, the three models with ERAS accuracies above 50 are also those with the smallest control–test differences for –/+ paradigms. This suggests that –/+ sentences are the ones for which our models are most likely to commit implicit garden path errors.

Word Boundary Supervision. The GPER results show that both whole word masking and joint training on word segmentation are effective methods for mitigating susceptibility to implicit garden path errors, and that using both techniques in combination results in the best GPER. However, the CWS and WWM+CWS models also have the poorest performance in terms of ERAS overall accuracy as well as ASAP test set accuracy. This suggests that joint training on word segmentation is effective in reducing susceptibility to garden path errors, but makes the model more susceptible to other types of errors, with the costs of the latter outweighing the benefits of the former. On the other hand, the

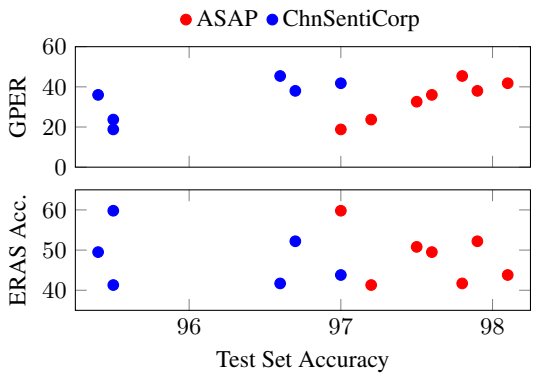

Figure 5: Test set accuracy is correlated with GPER, but exhibits no obvious relationship with ERAS accuracy.

use of whole word masking does not exhibit this tradeoff, though it only enables a modest reduction in GPER.

Test Accuracy and GPER. Figure 5 reveals a correlation between test set accuracy and GPER: better-performing sentiment analysis models are more susceptible to implicit garden path errors. No such relationship is observed between test set accuracy and ERAS overall accuracy, however. Thus, the increased susceptibility to garden path errors for better-performing models is offset by other factors that influence ERAS overall accuracy.

5 Discussion and Related Work

Left-Branching Bias in Humans. Much work has been done in psycholinguistics on the human processing of Chinese morphological garden paths (Inhoff and Wu, 2005; Huang and Li, 2020; Huang et al., 2021; Tong et al., 2023; Huang and Li, 2024). Huang and Li (2020), in particular, identify a human left-branching bias, similar to the left-branching bias observed for word segmentation models in Section 3. They find that human readers spend less time looking at left-branching morphological garden paths than right-branching ones, all else equal, indicating that the former are easier to process than the latter. On the other hand, Liao et al. (2024) did not observe a left-branching bias when instructing GPT-3.5 models to perform word segmentation.

Garden Paths in Word Segmentation. Rule-based, statistical, and non-neural machine learning methods have been proposed for correctly segmenting morphological garden paths in Chinese word segmentation (Sun and T'sou, 1995; Han et al., 2001; Li et al., 2003; Xiong and Zhu, 2007; Gao and Guo, 2009). These techniques mostly involve first identifying locally ambiguous substrings, possibly with the help of a lexicon of locally ambiguous substrings, and then choosing a segmentation for each of those substrings based on the surrounding context. Corpus analyses have found that morphological garden paths comprise roughly 4% of Chinese text (Qiao et al., 2008; Yen et al., 2012).

Targeted Evaluation in NLP. Minimal pair benchmarks have been used to evaluate language models' compliance with morphosyntactic constraints such as subject–verb agreement or negative polarity item licensing (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020). Fu et al. (2020) propose an error analysis framework that compares Chinese word segmentation models' performance across inputs with differing values for various features, such as sentence length or word frequency.

6 Conclusion

This paper has proposed a benchmark, ERAS, that detects garden path errors in Chinese NLP models, with or without an explicit word segmentation step in the pipeline. Using ERAS, we have observed that both word segmentation models and sentiment analysis models are vulnerable to garden path errors. In both cases, models seem to inherit this vulnerability from their training distributions. The MSR segmentation corpus most effectively defends against garden path errors, while the ASAP dataset seems to exacerbate susceptibility to garden path errors. We also find that garden path errors are driven by particular kinds of inputs: left-branching inputs in the case of word segmentation, and –/+ inputs in the case of sentiment analysis.

It has been argued that character-level tokenization is superior to word-level tokenization for Chinese neural NLP models (Li et al., 2019), and indeed, most Transformer language models for Chinese use character- or byte-level tokenization. Our findings from Section 4, however, have shown that a total lack of word boundary information makes a model highly vulnerable to implicit garden path errors. We have shown that injecting word boundary information into the model through whole word masking or joint training on word segmentation can significantly reduce implicit garden path errors, though care must be taken to avoid degradations in model performance. We leave the development of such measures to future work.

7 Limitations

ERAS can only be used for two kinds of models: word segmentation models and sentiment analysis models. In particular, our mode of evaluation for sentiment analysis models requires access to the model's logits, and therefore it is not compatible with large language models whose logits are not available to the user.

ERAS is also a large, synthetically generated dataset, consisting of minimal pair paradigms. Large synthetic minimal pair datasets are known to have the following limitations. First, large datasets are intrinsically impossible to evaluate fully (Bender et al., 2021), and synthetic minimal pair datasets are often found to contain examples that do not conform to the intended properties of the datasets (Blodgett et al., 2021). Second, minimal pair datasets have been criticized for lacking structural diversity (Javier Vazquez Martinez et al., 2023). ERAS is not an exception to either of these limitations.

8 Ethical Considerations

We are not aware of any ethical concerns arising from the work described in this paper.

References

Esma Balkir, Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2672–2686, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, OH, USA. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123–133, Jeju Island, South Korea. Association for Computational Linguistics.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. RethinkCWS: Is Chinese Word Segmentation a Solved Task? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5676–5686, Online. Association for Computational Linguistics.

Dongping Gao and Jiahong Guo. 2009. Dealing with Chinese Overlapping Ambiguity Based on Type Functional Application. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 3, pages 67–71, Shanghai, China. IEEE.

Dongli Han, Haodong Wu, and Teiji Furugori. 2001. Resolving overlapping ambiguities and selecting correct word sequence in chinese using internet corpus. *Journal of Natural Language Processing*, 8(3):107–121.

Linjieqiong Huang and Xingshan Li. 2020. Early, but not overwhelming: The effect of prior context on segmenting overlapping ambiguous strings when reading Chinese. *Quarterly Journal of Experimental Psychology*, 73(9):1382–1395.

Linjieqiong Huang and Xingshan Li. 2024. The effects of lexical- and sentence-level contextual cues on Chinese word segmentation. *Psychonomic Bulletin & Review*, 31(1):293–302.

Linjieqiong Huang, Adrian Staub, and Xingshan Li. 2021. Prior context influences lexical competition when segmenting Chinese overlapping ambiguous strings. *Journal of Memory and Language*, 118:104218.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020. Towards Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2062–2072, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Albrecht W. Inhoff and Caili Wu. 2005. Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & Cognition*, 33(8):1345–1356.

Hector Javier Vazquez Martinez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating Neural Language Models as Cognitive Models of Language Acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Zhen Ke, Liang Shi, Erli Meng, Bin Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Unified Multi-Criteria Chinese Word Segmentation with BERT. *Computing Research Repository*, arXiv:2004.05808.

Mu Li, Jianfeng Gao, Chang-Ning Huang, and Jianfeng Li. 2003. Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 1–7, Sapporo, Japan. Association for Computational Linguistics.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252, Florence, Italy. Association for Computational Linguistics.

Weiyan Liao, Zixuan Wang, Kathy Shum, Antoni B. Chan, and Janet Hsiao. 2024. Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 1961–1967, Rotterdam, Netherlands. eScholarship.

Chun Lin, Ying-Jia Lin, Chia-Jen Yeh, Yi-Ting Li, Ching-Wen Yang, and Hung-Yu Kao. 2023. Improving Multi-Criteria Chinese Word Segmentation through Learning Sentence Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12756–12763, Singapore. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4(0):521–535.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2022. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. *Computing Research Repository*, arXiv:1906.11455.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Wei Qiao, Maosong Sun, and Wolfgang Menzel. 2008. Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation. In *Text, Speech and Dialogue, 11th International Conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008, Proceedings*, pages 177–186, Brno, Czech Republic. Springer.

Maosong Sun and Benjamin K. T'sou. 1995. Ambiguity Resolution in Chinese Word Segmentation. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, pages 121–126, Hong Kong, China. City University of Hong Kong.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online. Association for Computational Linguistics.

Wen Tong, Jia Su, and Zhifang Liu. 2023. Competition in Overlapping Ambiguous Strings Segmentation: Does Word Frequency Have a Competitive Advantage Over Syntactic Prediction? *Social Science Research Network*, 4367593.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Shih-Ming Wang and Lun-Wei Ku. 2016. ANTUSD: A Large Chinese Sentiment Dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2697–2702, Portorož, Slovenia. European Language Resources Association (ELRA).

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ying Xiong and Jie Zhu. 2007. A New Machine Learning Method for Chinese Overlapping Disambiguity–Conditional Random Fields. In *2007 International Conference on Machine Learning and Cybernetics*, volume 7, pages 3922–3926, Hong Kong, China. IEEE.

Miao-Hsuan Yen, Ralph Radach, Ovid J.-L. Tzeng, and Jie-Li Tsai. 2012. Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and Writing*, 25(5):1007–1029.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833, Zurich, Switzerland. Springer International Publishing.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2023. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *Computing Research Repository*, arXiv:2209.02970.