# Correction of OCR Word Segmentation Errors in Articles from the ACL Collection through Neural Machine Translation Methods

**Vivi Nastase and Julian Hitschler**
University of Heidelberg
Heidelberg, Germany
{nastase, hitschler}@cl.uni-heidelberg.de

**Abstract**

Depending on the quality of the original document, Optical Character Recognition (OCR) can produce a range of errors – from erroneous letters to additional and spurious blank spaces. We applied a sequence-to-sequence machine translation system to correct word-segmentation OCR errors in scientific texts from the ACL collection with an estimated precision and recall above 0.95 on test data. We present the correction process and results.

**Keywords:** character-level sequence-to-sequence model, word segmentation, ACL collection

## 1. Introduction

The ACL anthology provides a valuable collection of scientific articles, and organizing it into a structured format could provide us with additional insight into research in this domain, help with finding related work and help with keeping up with new developments and ideas. The analysis of the ACL collection was stimulated by the shared task at ACL 2012 (Schäfer et al., 2012), the workshop on *Rediscovering 50 years of Discoveries*, and the series of SemEval tasks on *ScienceIE – Extracting Keyphrases and Relations from Scientific Publications*. Analysis has tackled various aspects of this collection: the citation network (Sim et al., 2012), citation references (Radev and Abu-Jbara, 2012; Gordon et al., 2016), keywords and relation extraction (Gábor et al., 2016), topics and community studies (Bordea et al., 2014), among others.

Considering the status of our NLP toolbox, the success on processing such a corpus increases when the texts are clean. The ACL collection contains numerous articles published before electronic submission became standard for our conferences. These older papers have been scanned and processed through OCR, resulting in texts that contain errors. A brief inspection of this portion of the collection has shown that a very common error introduced by the OCR process consists of spurious blank spaces, which split words randomly in a varying number of smaller fragments (Figure 1). This problem is so pervasive, that it influences subsequent processing, for example keyword extraction – in the list of keywords produced with the SAFFRON system, we found the following keywords in papers from Coling 1996: *Non-l inear*, *Context fi*, *Segmentat ion*, *Ion theory*.

Inspired by previous work on error correction using machine translation models (e.g. (Yannakoudakis et al., 2017)), we apply a character-level sequence-to-sequence model to learn how to segment English words in the ACL collection. Written modern languages of European origin usually segment words explicitly, so English texts generally do not require word segmentation, but as we have seen, this problem has popped up in documents processed with OCR. This means that we can very easily obtain large volumes of data for training. We use a portion of the ACL collection (the articles published after electronic submission had become the norm in our community, a date which we conservatively set at 2005) to generate training, development and test data. The high results – above 96% precision – obtained on the test data indicate that processing the part of the collection published before 2005 would solve the vast majority of the word fragmentation issues, and would provide the community with a corpus of increased quality. In this paper, we present the processing tool and experiments done on the post-2005 portion of the collection, and the corrected ACL collection will be offered to the ACL anthology editor to be made available to the community.[1]

## 2. Related Work

Depending on their source, errors in unedited texts can fall into various categories: typos, deliberate misspellings including shortened/phonetically written words (particuarly on social media), word segmentation errors, erroneous characters, non-canonical spellings (historical texts), grammatical errors, and probably more.

For many of the above mentioned problems, neural-based approaches originally developed for machine translation have proved to be very successful. Yannakoudakis et al. (2017) use a machine translation inpired approach – N-best list ranking using neural sequence labelling models – for grammatical error correction. Word and character-based sequence-to-sequence models (Yuan and Briscoe, 2016; Xie et al., 2016; Yang et al., 2017) have achieved good performance on the CoNLL-2014 shared task (Ng et al., 2014) on text correction (which covers a variety of errors made in written essays by second language learners). Schmaltz et al. (2017) show that character-level sequence-to-sequence models perform better than word-level models even with less training data than previous sequence-to-sequence ap-

[1] Information about how the word re-segmented version of the ACL corpus is available will be posted at `http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/ACL_corrected/ACL_corrected.shtml`.

```
INTRODUCTION
Although the l i terature dealing with formal and natural
languages abounds with theoretical arguments of worst-
case performance by various parsing strategies \[e.g.,
Griffiths & Petrick, 1965; Aho & Ullman, 1972; Graham,
Harrison & Ruzzo, Ig80\], there is l i t t le  discussion of
comparative performance based on actual practice in
understanding natural language. Yet important practical
considerations do arise when writing programs to under-
stand one aspect or another of natural language utteran-
ces. Where, for example, a theorist wi l l  characterize a
parsing strategy according to its space and/or time
requirements in attempting to analyze the worst possible
input acc3rding to ˜n arbitrary grammar st r ic t ly  limited
in expressive power, the researcher studying Natural
Language Processing can be just i f ied in concerning
himself more with issues of practical performance in
parsing sentences encountered in language as humans
Actually use i t  using a grammar expressed in a form
corve˜ie: to the human linguist who is writing i t .
```

Figure 1: Excerpt from paper P81-1001, *A Practical Comparison of Parsing Strategies* by Jonathan Slocum

proaches, and outperform statistical phrase-based machine translation models on the CoNLL data.

Chen et al. (2015a; Chen et al. (2015b) explore the use of GRUs and LSTMs for Chinese word segmentation, and Zhang et al. (2016) approach the task using word and character context in a globally optimized beam-search framework for neural structured prediction. Yang et al. (2017) build on this previous work to produce a modular neural-based segmentation model for Chinese, using five-character window, pre-trained based on a variety of external resources.

Based on these previous analyses into the kind of architectures that perform well for different types of error correction, we adopt a character-level sequence-to-sequence model for the word segmentation of English texts.

## 3. The ACL collection

The ACL collection we work with consists of 18,849 scientific articles published between 1965 and 2012. Papers published before electronic submission became the norm in the 2000s have been scanned and processed with OCR, and as such suffer from the type of errors common in such material – most notably spurious spaces and erroneous characters.

The most common OCR error we noticed was incorrect word segmentation – there are numerous spurious blank characters at random locations in the texts, as can be seen in the text fragment from paper P81-1001 displayed in Figure 1 (which we reproduce as is in the file, including new lines). The problem is so pervasive that it influences tasks such as keyword extraction, the basis for further processing of the collection. This is evidenced by an inspection of the keywords produced with the SAFFRON system (Bordea et al., 2014), wherein we find keywords among the top 15 ranked for each articles that were affected by this OCR error. While wrongly split keywords may not seem so problematic as long as all parts are present (e.g. *Segmentat ion*), incomplete words may lead to erroneous or misleading keyphrases (e.g. *Context fi* and *Ion theory*). We set to address this problem, considering that training and test data can be automatically obtained from the portion of the ACL collection that consist of electronic submissions (conservatively, we choose 2005 as our lower time limit).

## 4. Machine translation-based correction model

We use the *nematus* system[2] (Sennrich et al., 2017), a state-of-the-art sequence-to-sequence machine translation model. It is a highly configurable system that implements an attentional encoder-decoder architecture. For the experiments presented here we use the default cross-entropy minimization as the training objective, via (accelerated) stochastic gradient descent. We use this system to process sequences at the character level. The training data consists of parallel input-output sequences, with a default limit of sequence length 100. Below we describe what kind of training data was provided to the system.

As noted from the fragment in Figure 1, the scanned text preserves the line breaks from the original paper, which include hyphenated words. The first processing step we apply to the entire collection is to remove the new lines if the line does not finish with a dot, question mark or colon. Hyphenated words are replaced with their non-hyphenated version if such a variant was encountered anywhere in the texts. This processing step produces texts with one paragraph per line. After this step, we separated the collection – pre-2005 (B2005) (to be conservative about the beginning of widespread use of OCR) and post-/ including 2005 (A2005). The texts from A2005 were considered correct from the point of view of word segmentation, and the training data was produced from these texts, taking into account

[2] https://github.com/EdinburghNLP/nematus

| input | output |
|---|---|
| t o - i n f i n i t i v e s | t o - i n f i n i t i v e s |
| a n d g e r u n d s | a n d ## g e r u n d s |
| B o t h t h e b a s e l i n e a n d S p a d e o p e r a t e o n p a r s e | B o t h ## t h e ## b a s e l i n e ## a n d ## S p a d e ## o p e r a t e ## o n ## p a r s e |
| t r e e s w h i c h w e r e o b t a i n e d f r o m C h a r n i a k ? s | t r e e s ## w h i c h ## w e r e ## o b t a i n e d ## f r o m ## C h a r n i a k ? s |
| p a r s e r | p a r s e r |
| O u r s e t o f e x p e r i m e n t a l m a t e r i a l s c o n t a i n e d | O u r ## s e t ## o f ## e x p e r i m e n t a l ## m a t e r i a l s ## c o n t a i n e d |
| c o m p r e s s i o n s | c o m p r e s s i o n s |
| P r o c e d u r e a n d S u b j e c t s W e o b t a i n e d c o m p r e s s i o n | P r o c e d u r e ## a n d ## S u b j e c t s ## W e ## o b t a i n e d ## c o m p r e s s i o n |
| r a t i n g s d u r i n g a n e l i c i t a t i o n s t u d y c o m p l e t e d | r a t i n g s ## d u r i n g ## a n ## e l i c i t a t i o n ## s t u d y ## c o m p l e t e d |

Table 1: Example of input-output parallel training data for correcting word segmentation problems in the ACL collection. The ## sequence indicates a blank space.

| |
|---|
| predicted<br>`001) and by items ( F2(3; 117) = 40 \/\/ 0 0 1 )` |
| gold standard<br>`001) and by items ( F2(3;117) = 40 \/\/ 0 0 1 )` |

where ”F2(3;117)” is considered one word in the gold standard, and is split in two in the predicted version.

Table 2: Errors caused by erroneous spacing in formulas.

the chosen sequence limit, as follows:

1. the texts with one paragraph per line are split into smaller fragments, avoiding as much as possible splitting on ”ambiguous” breaking points (i.e. spaces between text fragments which may actually be erroneous):
   (a) split on end of sentence characters or phrase delimiting characters (.?!;: - parentheses)
   (b) if the fragment is longer than 50 characters, split at numbers
   (c) if the fragment is still longer than 50 characters, split into 50 character long sequences
2. produce nematus input training data by removing blank spaces from the string, and (conform nematus' input formatting) inserting a blank space after each character
3. produce nematus output training data by replacing blank spaces with a special sequence (##) and then inserting a blank space after each character (except the ones in the special sequence).

The parallel input-output training data is exemplified in Table 1.
The data prepared in this manner consists of 9,310,664 input-output parallel sequences. We selected 2,000,000 sequences for training, and 500,000 sequences for testing. Training was done using the system's default settings – training is done with cross-entropy minimization with adam optimizer, encoder and decoder implement GRUs, learning rate 0.0001, embedding layer size 512, hidden layer size 1000, dropout for input embeddings and hidden layers 0.2. The model built during training is used to ”translate” the input test data, which are then compared token-by-token to the expected output test sequences.

## 5. Results and discussion

We have performed two evaluations: one with respect to the test data described in Section 3., and one on the actually corrected data, the B2005 portion. The results of these evaluations are described in Sections 5.1. and 5.2. respectively.

### 5.1. Evaluation on test data

For the ACL correction, evaluation was performed on 500,000 fragments obtained as explained in Section 3.. We evaluate in terms of word-level precision and recall, computing the number of correctly predicted words. Formally, for an automatically produced sequence $\mathbf{w}_a$, we compute precision and recall by comparison with a gold standard sequence $\mathbf{w}_{gs}$[3]:

[3]We consider these sequences as ordered lists of words, and in evaluation we gradually shorten the list such that two occurrences of the same word in the automatically produced output are compared to different tokens in the gold standard.

```
Although the literature dealing with formal and natural languages abounds
with theoretical arguments of worst-case performance by various parsing
strategies \[ e  . g . , Griffiths & Petrick , 1965; Aho & Ullman , 1972;
Graham , Harrison & Ruzzo , Ig 80\] , there is little discussion of comparative
performance based on actual practice in understanding natural language  . Yet
important practical considerations do arise when writing programs to understand
one aspect or another of natural language utteran-ces  . Where , for example ,
a theorist will characterize a parsing strategy according to its space and/or
time requirements in attempting to analyze the worst possible input  acc3rding
to ˜ n arbitrary grammar strictly limited in expressive power  , the researcher
studying Natural Language Processing can be justified in concerning himself
more with issues of practical performance in parsing sentences encountered in
language as humans Actually use it using a grammar expressed in a form corve ˜
ie : to the human linguist who is writing it  .
```

Figure 2: Corrected version of the excerpt from paper P81-1001, *A Practical Comparison of Parsing Strategies* by Jonathan Slocum from Figure 1

$$Prec = \frac{|\{w|w \in \mathbf{w}_{gs}, w \in \mathbf{w}_a\}|}{|\mathbf{w}_a|}$$

$$Rec = \frac{|\{w|w \in \mathbf{w}_{gs}, w \in \mathbf{w}_a\}|}{|\mathbf{w}_{gs}|}$$

Precision and recall on the entire test data is a micro-average of the scores for the 500,000 sequences in the test data. We obtained a precision of 0.955 and recall of 0.950 on re-segmenting into words. Most of the errors we observed are caused by spaces in formulas, as in the examples in Table 2. Partially discounting this type of error (which impacts very little, if at all the text processing of the ACL collection), the precision and recall become 0.979 and 0.974 respectively.

Because we want to apply the trained model to the B2005 portion of the collection for which we have no test data, we would like to have a better idea of what the results are likely to be. The fact that the papers in the B2005 portion of the collection come from the same domain as the A2005 portion which was used for training makes it highly likely that most of the vocabulary in B2005 is shared with A2005, but there will also be tokens unknown to the model. For this reason we performed an additional evaluation on the A2005 test data, for tokens that do not appear in the training or development data. The recall on these unknown tokens is 0.821, and the precision 0.876. Many of these unknown tokens are part of formulas, which, as mentioned before, we think is highly unlikely to impact information/keyword extraction, and other such tasks that focus on "proper" text.

## 5.2. Evaluation on corrected data (B2005)

The high performance on the test data and on unknown tokens indicates that applying the model to the B2005 collection will likely solve many of the existing segmentation problems. In Figure 2 we include the version obtained after the word re-segmentation process of the fragment in Figure 1.

The motivation that lead us to attempt correcting word segmentation errors was that they were so prevalent that even keywords presented such problems. Within the used ACL anthology was a file of keywords obtained with the SAFFRON system (Bordea et al., 2014). This version of the system is no longer available to obtain keywords on the corrected version of the files. Because of this we decided to perform an evaluation with respect to the keywords obtained on the raw files: we selected randomly 40 papers from the collection published before 2005, and annotated each of their keywords as *correct* or *incorrect* with respect to word segmentation[4]. This provided a total of 573 keywords, 38 of which were incorrect, and 535 correct. We tested each of these keywords against the raw and corrected versions of the corresponding files, and present the summary of results in Table 3.

| tag | N | in the raw text | in the corrected text |
|---|---|---|---|
| incorrect | 38 | 35 (92.1%) | 3 (7.9%) |
| correct | 535 | 455 (85.05%) | 453 (84.67%) |

Table 3: Results on the keywords from a sample of 40 documents

Of the 38 incorrect keywords, 35 actually appear in the raw files. Those that do not appear in the raw files seem to be caused by some preprocessing done by the keyword extractor – e.g. for the paper A94-1014, the keyword that does not appear in the raw (or corrected) file is "Computer Science Univ", which seems to have been caused by collapsing the lines that contain the authors' affiliations ("Dept . of Computer Science [new line] Univ . of Central Florida"). Only 3 of the 38 badly segmented keywords appear in the corrected files.

For the correct keywords, the reason why some do not appear in the raw or the corrected files is mainly the preprocessing done by the keyword extractor (e.g. lemmatization). There are two correct keywords that appear in the raw files but not in the corrected files. One of these is *language process* (from paper P84-1101), which seems correct

[4]The annotation was performed by one of the authors of the paper. The problem is quite obvious, and there didn't seem to be a need for an additional annotator and agreement computations.

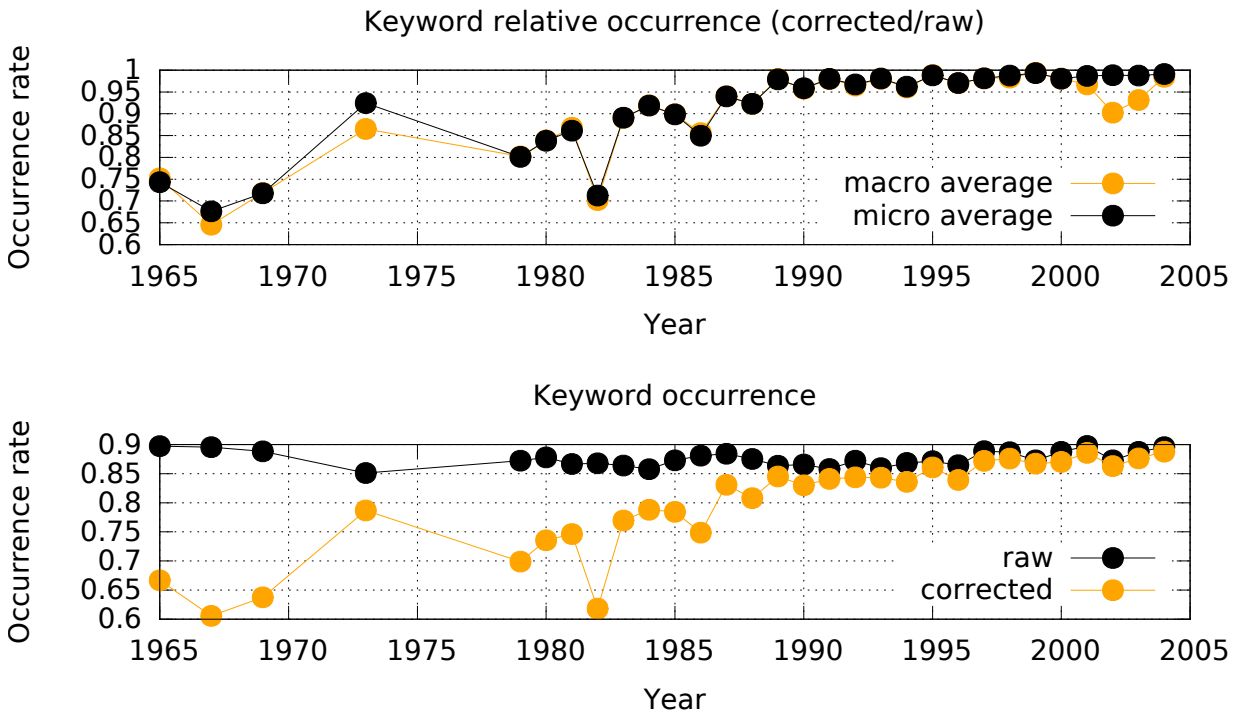


Figure 3: Proportion of keywords from the original (raw) files in the raw files and the corrected versions.

but in fact was extracted because of the erroneous segmentation *language process ing*. The second error is a proper processing error: the keyphrase *rigid system-directed dialogue structure* (from paper E93-1061) appears in the corrected version as *arigid system-directed dialogue structure*, "rigid" having been merged with the preceding indefinite article.

The results of the manual analysis show that the sequence to sequence model does indeed perform well on correcting the B2005. We performed an additional analysis to obtain another estimation of the impact of the word re-segmentation. A visual inspection of the ACL data shows that older papers appear more difficult for OCR processing as they seem to have been first scanned from prints of various quality. More recent papers were more often processed from directly produced pdf files. This would indicate that older papers have more OCR (and thus, word segmentation) errors compared to more recent ones. This would mean that more recent papers have more of the extracted keywords (because they would be more likely to be correct) than older papers. We plot the rates of occurrence of the keywords in the raw and corrected versions on the files, by year, for the B2005 articles that have keywords. The top of Figure 3 shows the macro and micro average of the number of keywords that appear in the corrected vs. the raw articles (number of keywords that appear in the corrected files divided by the number of keywords in the raw files). The bottom figure plots separately the rate of occurrence of the keywords in the raw and corrected files respectively, relative to the total number of keywords for each paper. The upward trend of the ratio of old keywords in the corrected files confirms the above observation, as in older papers fewer keywords appear in the corrected versions of the papers than in more recent ones, correlating with the increase in OCR quality.

The B2005 portion of the ACL collection was processed, and the re-segmented texts will be offered to the ACL anthology editors. Information about availability will be posted on the website of the University of Heidelberg's Computational Linguistics Institute[5].

## 6. Conclusion

We have presented the character-level sequence-to-sequence model used to correct one of the very pervasive errors in the part of the ACL collection processed through OCR – spurious blank spaces that fragment words. The high results on the test portion of the data indicate that a large part of this type of errors could be corrected in the ACL collection. We have applied this process and produced a cleaner version of the ACL collection, which we will offer to the ACL anthology editors to make it available with the raw collection to the community.

## 7. Acknowledgements

We thank our reviewers for their comments. This research was funded by the Leibniz Science Campus Empirical Linguistics & Computational Language Modeling, supported by Leibniz Association grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art of Baden-Württemberg.

## 8. Bibliographical References

Bordea, G., Buitelaar, P., and Coughlan, B. (2014). Hot topics and schisms in NLP: Community and trend anal-

[5] http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/ACL_corrected/ACL_corrected.shtml

ysis with Saffron on ACL and LREC proceedings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Chen, X., Qiu, X., Zhu, C., and Huang, X. (2015a). Gated recursive neural network for chinese word segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1744–1753, Beijing, China, July. Association for Computational Linguistics.

Chen, X., Qiu, X., Zhu, C., Liu, P., and Huang, X. (2015b). Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, Lisbon, Portugal, September. Association for Computational Linguistics.

Gábor, K., Zargayouna, H., Buscaldi, D., Tellier, I., and Charnois, T. (2016). Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *LREC 2016*.

Gordon, J., Zhu, L., Galstyan, A., Natarajan, P., and Burns, G. (2016). Modeling concept dependencies in a scientific corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–875. Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June. Association for Computational Linguistics.

Radev, D. and Abu-Jbara, A. (2012). Rediscovering acl discoveries through the lens of acl anthology network citing sentences. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 1–12. Association for Computational Linguistics.

Schäfer, U., Read, J., and Oepen, S. (2012). Towards an acl anthology corpus with logical document structure: an overview of the acl 2012 contributed task. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 88–97. Association for Computational Linguistics.

Schmaltz, A., Kim, Y., Rush, A., and Shieber, S. (2017). Adapting sequence models for sentence correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2797–2803, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Sim, Y., Smith, N. A., and Smith, D. A. (2012). Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 22–32. Association for Computational Linguistics.

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., and Ng, A. Y. (2016). Neural language correction with character-based attention. *CoRR*, abs/1603.09727.

Yang, J., Zhang, Y., and Dong, F. (2017). Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Vancouver, Canada, July. Association for Computational Linguistics.

Yannakoudakis, H., Rei, M., Andersen, Ø. E., and Yuan, Z. (2017). Neural sequence-labelling models for grammatical error correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2796, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yuan, Z. and Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California, June. Association for Computational Linguistics.

Zhang, M., Zhang, Y., and Fu, G. (2016). Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 421–431, Berlin, Germany, August. Association for Computational Linguistics.