Transformer-Based Real-Word Spelling Error Feedback with Configurable Confusion Sets

Torsten Zesch and **Dominic Gardner** and **Marie Bexte**
CATALPA – Center of Advanced Technology for Assisted Learning and Predictive Analytics
FernUniversität in Hagen, Germany
Correspondence: torsten.zesch@fernuni-hagen.de

Abstract

Real-word spelling errors (RWSEs) pose special challenges for detection methods, as they 'hide' in the form of another existing word and in many cases even fit in syntactically. We present a modern Transformer-based implementation of earlier probabilistic methods based on confusion sets and show that RWSEs can be detected with a good balance between missing errors and raising too many false alarms. The confusion sets are dynamically configurable, allowing teachers to easily adjust which errors trigger feedback.

1 Introduction

Real-word spelling errors (RWSE) are specific spelling mistakes, where the resulting misspelling is another existing word:

> *Time flies like an error [arrow].*[1]

This is in contrast to non-word spelling errors, where the resulting string is out-of-vocabulary:

> *Time flies like an arro [arrow].*

The distinction is grounded in lexical inclusion criteria for a given language, a problem that is itself non-trivial. In this paper, we consider the lexicon of a language to be provided as a fixed list containing not only lemmas, but also inflected forms. The list does not contain open classes like complex named entities or other noun compounds.

Of special interest are RWSEs where the sentence with the error is syntactically well-formed, so that they can only be detected when taking semantic information into account. Compare the following two examples:

1. *The name comes from the Greek work [word] for sun.*
2. *These plants are more tolerance [tolerant] to drought.*

In the first example, the RWSE is not readily detectable via syntactic analysis, whereas in the second example syntax alone provides some evidence for a possible error.

RWSEs are quite common in English, but also happen in other languages e.g. in German:

> *Er ist eine Konifere [Koryphäe] auf seinem Gebiet.*
> *(He is a conifer [coryphaeus] in his area.)*

This specific example is sometimes deliberately inserted for comical effect as the contrast of 'big expert' with 'small tree' can be considered funny. However, there are other, less pretentious examples, like replacing *art* or *part* with *fart* that can be quite embarrassing if unintentional.[2]

While there is a long tail of idiosyncratic RWSEs, some are also quite common and can be considered **confusion sets** (Golding and Schabes, 1996), i.e. fixed sets of words that are often confused with each other – especially by language learners. Examples include *{dessert,desert}*, *{peace,piece}*, *{sight,site}*, *{than,then}*, or *{their,there}*. Note that the sets are not ordered, so that e.g. *sight* could be inserted for *site* or vice versa.

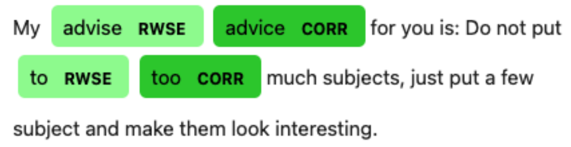
The ability to detect RWSEs reliably is essential for enabling automated feedback on this class of errors. In this paper, we review the related work and find that a modern implementation for finding RWSEs is missing. We thus propose a Transformer-based approach with configurable confusion sets, which will give teachers the ability to select which words are currently in focus, so that targeted feedback can be provided. Figure 1 gives an example.

We make the RWSE-checker available as an open-source implementation together with a demo

[1] When giving examples, we always put the error first and the [correction] in square brackets. When referencing a confusion set outside of an example, we use {token1, token2}.

[2] The subset of similar sounding RWSEs that are often used for comical effect is also called *malapropism*. An unexpectedly fitting or creative malapropism is also called *eggcorn* (itself an eggcorn of 'acorn'). Eggcorns are often coined by language learners trying to make sense of an unfamiliar word or phrase that they have not yet seen in writing. A famous example is 'old-timers' disease' for 'Alzheimer's disease'.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 375–383
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics



My advise RWSE advice CORR for you is: Do not put to RWSE too CORR much subjects, just put a few subject and make them look interesting.

Figure 1: Example for highlighting RWSEs as part of writing feedback.

application and all our experimental code.[3]

2 Related Work

Early approaches to RWSE correction either relied on measuring the local contextual fitness of words through semantic-relatedness measures (Budanitsky and Hirst, 2006) or n-gram language models (Mays et al., 1991; Wilcox-O'Hearn et al., 2008), where after detecting a word with low contextual fitness a neighborhood space of candidate replacements was searched for a better fitting one. While such approaches are flexible and can find all kinds of RWSEs, they are computationally costly (as they have to test for each word in a sentence a potentially large number of candidates), and yield a lot of false alarms as they often detect 'errors' that are e.g. synonyms of the original word.

As a way around those challenges, other early approaches relied on the already introduced confusion sets, i.e. they limited the search to known target words and a very small set of candidates.[4] At the same time, before the availability of large language models, it was much faster to train a supervised classifier for each confusion set (Golding and Schabes, 1996; Carlson et al., 2001).

Another related field is *Grammatical Error Correction* (GEC), i.e. the process of detecting and correcting grammatical errors in text (Ng et al., 2013; Yuan and Briscoe, 2016). Most recent approaches use a seq2seq design where the text with errors is transformed into an error-free version. In doing so, a GEC system might also fix RWSEs along the way, but as it targets all kinds of errors, we might not know where a RWSE occurred which limits the kind of feedback we can give. Error types are only considered post-hoc and common schemes do not distinguish between non-word and real-word spelling errors (Bryant et al., 2017).

So our approach combines ideas from earlier work: (i) we rely on confusion sets, but without the supervised classifiers, making the sets dynamically configurable); we find RWSEs with the help of language models, but using masked language models and limiting the candidate space through confusion sets.

Confusion sets have also been used in unsupervised GEC approaches to generate candidate sentences that are then scored by a Transformer-based model (Bryant and Briscoe, 2018; Alikaniotis and Raheja, 2019). Our approach can be seen as a special case, where we only use RWSE confusion sets.

Technically, finding RWSEs in such a way is similar to *lexical substitution* (Zhou et al., 2019), with the crucial difference that an RWSE is an implausible word that is substituted with a more plausible one, while in lexical substitutions both words need to be plausible in the given context.

3 Method

Our implementation is based on the *fill-mask* task[5] of the Transformer library. Given a sentence like

> *People with lots of honey usually live in big houses.*

a word is replaced with a mask token and the library returns the most likely fillers and their probabilities. So for the resulting masked sentence

> *People with lots of [MASK] usually live in big houses.*

we get the following results:

```
   money: 0.522
  wealth: 0.053
children: 0.022
  income: 0.016
  family: 0.014
```

The original token *honey* is not even in the top-5 and *money* is one order of magnitude more likely than the next candidate.

However, as we do not know where to look for errors (remember that RWSEs are in-vocabulary and thus hard to detect), we would have to test every token in a sentence which would be quite costly. Also, even if we are ready to invest the compute, blindly following this approach could introduce new errors. For example,

> *People with lots of money usually live in big [MASK].*

[3] https://github.com/zesch/rwse-experiments

[4] Some approaches allow the empty word in confusion sets to cover also insertions or deletions, but most papers (and we in our study) limit confusion sets to replacements.

[5] https://huggingface.co/tasks/fill-mask

returns *cities* with a probability of 0.74 and would thus result in a false alarm. We thus combine this approach with **confusion sets**. In our example, we would only test for {money, honey} and get the following result:

```
money: 0.52241
honey: 0.00004
```

Note that our example is for illustrative purposes, but that in a real setting with pedagogically relevant confusions {money, honey} would probably not be a target confusion set.

Threshold Factor While in the {money, honey} example above, the correct choice was several orders of magnitude more likely than the mistake, this might not always be the case. Especially words with a high prior probability might lead to false alarms. We thus introduce the magnitude parameter μ indicating how many times more likely a candidate needs to be in order to be considered as a replacement. We initially set $\mu = 10$ so that a RWSE candidate needs to be an order of magnitude more likely, but will later more formally analyze the impact of this parameter, similar to the analysis in Carlson et al. (2001).

4 Experimental Setup

There are currently more than 14,000 models on Hugging Face that are compatible with the fill-mask task. As we are mainly conducting experiments with English text and are interested in production-grade performance, we stick with the basic `google/bert-base-cased` Transformer model.[6]

4.1 Confusion Sets

We compiled a list of pedagogically relevant confusion sets by scanning prior work (Golding and Schabes, 1996; Carlson et al., 2001), but also consulting with domain experts. As a limitation of the fill-mask task is that it cannot directly return probabilities for words that are not in the model vocabulary, we discard confusion sets where at least one element of the set is out-of-vocabulary. Another limitation of the fill-mask task is that it only works with single tokens. So we also discard the few cases where multi-word tokens are involved, e.g. {a life, alive}. We also discard confusion sets with apostrophes like {its, it's}. Our final list contains 52 confusion sets. Table 1 gives an overview.

As capitalization can be an important source of information, we work with a cased BERT model and differentiate between lower case and upper case variants of each token. Thus, while Table 1 only lists lower-case forms for better readability, a confusion set also usually contains the upper-case variants. We mark this with an underlined first letter. A missing underline indicates that the upper-case form was not in the vocabulary and was thus discarded (as it could not be predicted anyway and would raise an error message).[7]

4.2 Data

As datasets of naturally occurring RWSEs are extremely rare, we mainly focus on synthetic datasets.

We use a news sentence base, as we expect to find very few naturally-occurring RWSEs (which would distort our experiments) in the professionally edited news texts. We select from the Leipzig Corpora Collection (Goldhahn et al., 2012) the NEWS dataset with 10,000 English sentences from 2023. If a token in a sentence matches an entry in one of our confusion sets, the sentence is retained. Out of 10,000 sentences 7,344 contain at least one of our confusion sets. Many sentences trigger more than one. Table 1 shows how often each confusion set was found overall.

While evaluating on this dataset will provide a realistic impression of the expected performance, we cannot analyze which confusion sets are most challenging as many appear too infrequently. We thus create another dataset (called NEWS-BALANCED) by randomly iterating over a the largest download from the Leipzig Corpora Collection with 1 million sentences. Whenever a confusion set is triggered, we keep the sentence up to a maximum of 100 instances. However, even within 1 million sentences, some confusion sets do not appear 100 times, e.g. only 11 sentences were found for 'Provence'. Instead of sampling from an even larger sentence base, we accept the slight imprecision. Results obtained that way are still valid and interpretable, only a bit less reliable. In the balanced dataset, the extracted sentences were used with the intended confusion set only, so as not to trigger multiple RWSEs which would distort results (e.g. the confusion set {to, too, two} would be triggered in almost all sentences in addition to the actually sampled

[6] https://huggingface.co/google-bert/bert-base-cased

[7] To give an example: *begin / being* should be interpreted as 'begin' can be confused with either 'being' or 'Being'.

{to, too, two}	6,034	{life, live}	166	{forth, fourth}	37		
{their, there, they}	1,860	{mad, made}	157	{raise, rise}	36	{extend, extent}	16
{width, with}	1,556	{country, county}	145	{hole, whole}	34	{principal, principle}	13
{you, your}	998	{weak, week}	131	{trail, trial}	34	{plain, plane}	12
{form, from}	919	{found, fund}	129	{ease, easy}	32	{Provence, province}	10
{were, where}	650	{few, view}	109	{affect, effect}	32	{spit, split}	9
{or, ore}	454	{lead, led}	97	{peace, piece}	32	{desert, dessert}	7
{than, then}	446	{passed, past}	81	{sight, site}	30	{brakes, breaks}	7
{which, witch}	443	{things, thinks}	70	{affects, effects}	23	{bitch, pitch}	7
{them, theme}	261	{weather, whether}	66	{feat, feet}	22	{forms, forums}	3
{begin, being}	231	{safe, save}	53	{accept, except}	21	{crab, crap}	2
{three, tree}	184	{capital, Capitol}	52	{advice, advise}	21	{weed, wheat}	1
{word, world}	167	{quiet, quite}	39	{loose, lose}	20		

Table 1: List of confusion sets and their frequency in the NEWS dataset

set). With these two sentences bases, we now introduce synthetic errors by replacing the token from the confusion set found in the sentence with each of the other words in the set.

As synthetic datasets are probably underestimating the difficulty of the task, we are using one of the few available datasets of naturally occurring RWSEs from Zesch (2012) which was created by mining the Wikipedia revision history. The dataset is quite small and some of their confusion sets do not match our RWSE definition (e.g. containing singular/plural of the same noun). We decided to also use the German version of the dataset in addition to the English one. We manually cleaned both versions and arrive at 49 English sentences (WIKI-EN) and 30 German sentences (WIKI-DE).[8] Finally, we are using as EDUCATIONAL texts the 1,244 exam scripts from the CLC-FCE corpus (Yannakoudakis et al., 2011) from which we extract 18,984 sentences.

4.3 Evaluation Metrics

Model performance was evaluated based on two different classification metrics: *false-alarm rate* (or false positive rate) and *miss rate* (or false negative rate).

False-alarm rate is computed as the ratio of false-alarms to all ground truth negatives which are equal to all detection instances assuming that the dataset is error-free. In the NEWS datasets, we just take the original sentences, which we consider to be almost RWSE-free. In the WIKI datasets, we know all RWSE instances and create the corrected versions. In the EDUCATIONAL dataset, we have no way of knowing a-priori where to find RWSEs, so we manually evaluate all triggered detection instances to determine real false alarms.

Regarding missed RWSE detections, we take the synthetic part of the NEWS dataset, i.e. the NEWS sentences where we introduced mistakes, and compute the miss rate as the number of instances where the original word is not selected divided by the total number of all synthetic instances. In the WIKI dataset, we do the same with the naturally occurring errors. In the EDUCATIONAL dataset, we cannot easily compute miss rate as this would require a full normalization.

5 Results & Discussion

Table 2 provides an overview of the high-level results. We generally see very low *false-alarm* and *miss rates* on the NEWS datasets. False alarms and misses increase by one order of magnitude on the datasets of documented RWSEs from WIKIpedia. Also on the EDUCATIONAL dataset, the false-alarm rate increases about 100-fold compared to the NEWS dataset (from .001 to .107). However, we still consider such false alarm rates acceptable as about 90% of RWSEs are correctly identified, for which we then can provide feedback.

It is hard to compare our results with previous results, as the originally used synthetic datasets are not available, different confusion sets were used, and results depend much on parameter choice (especially our magnitude parameter μ). Carlson et al. (2001) report results on synthetic data for different prediction thresholds that serve a similar purpose as our parameter μ. They report a 'performance' (which we interpret as accuracy) of .981 for 19 highly frequent confusion sets, and .973 for a larger set of 265 confusion sets. Converting our metrics into accuracy, we obtain on the NEWS dataset an accuracy of .965 (for our 52 confusion sets and $\mu = 10$), which is in the same ballpark but as discussed above not directly comparable.

[8] For German language texts, we used the multilingual variant `google/bert-base-multilingual-cased` and the confusion sets defined by Zesch (2012).

Dataset	# no RWSE	false alarm rate	# actual RWSEs	miss rate
NEWS	15,960	.001	60,632	.005
NEWS-BALANCED	15,454	.011	40,400	.044
WIKI	49	.082	49	.163
WIKI-DE	30	.000	30	.100
EDUCATIONAL	-	.105	-	-

Table 2: Overall RWSE detection results with $\mu = 10$.

5.1 Qualitative Analysis on NEWS

The 10,000 sentences trigger 15,960 confusion sets (see Table 1 for the distribution) in the NEWS dataset. Our RWSE detection method produces only 13 false alarms which we further analyze here. Out of the 13 false alarms, we only consider three to be clear-cut mistakes. Six false alarms can be attributed to contextual ambiguity that allows for both, the original and suggested token, to be applicable. The remaining mistakes can be blamed on incomplete sentences or actual errors in the sentences.

{country,county} This confusion set produces 3 related false alarms, e.g. *An Officer of the OBE is awarded for distinguished regional or [county]-wide role in any field, through achievement or service to the community.*
In those cases both words of the confusion set appear plausible without knowing the broader context.

{form,from} *Crowds were entertained by the Broke FMX Motocross Stunt Team, as well as a crowd pleasing display [form] the Pony club Games.*
This is probably a RWSE in the original sentence and should not count as false alarm.

{hole,whole} *That's the [whole] in the end: A single guess provides you with information that you then need to use to narrow down the list of subsequent guesses.*
This sentence seems incomplete. Adding 'story' after the token could resolve this and 'whole' would be correct.

{life,live} *In 2014 they moved to Mauriceville TX. where they built a beautiful home and [life] together.*
Without a wider context, 'live' would also be plausible.

{their,there,they} *[They] were suspected bodies of soldiers killed in a then recent attack on the Melete barracks.*
Within the wider context of this sentence[9] 'They' seems appropriate, so we do not count this as a clear error.

Many open from 8 or 9am but you can refer to [their] to confirm if your local restaurant is open during the Christmas period and what times they are trading.
The model suggests 'there', which seems equally wrong as the original token. So we do not treat this a clear false-alarm.

But [they] are two distinct and separate occurrences.
The model suggests 'there', but without wider context both versions are acceptable. So we do not treat this as a clear false-alarm.

{theme,them} *Now we get to the timeliness — and Novey's knack for being on [theme] without ever being too on the nose.*
This is an actual false alarm.

{to,too,two} *She also helped lead the girls basketball and softball teams [two] section championships.*
The model predicts `to` with high score of 0.97, which is not clearly better. Another likely solution would haven been 'to two section championships' instead.

{you,your} *Next thing [your] going to post is that chopping the foreskin of babies isn't a symbol of virtuous enlightenment.*
This should likely have been `you're`, but the current implementation does not support contractions. As this would lead to wrong feedback, we count this as a false alarm.

Probably not a real good look for [you] trade demand that you're posting a pic excited with OBJ being a Raven.
This is clearly a true false alarm.

5.2 Confusion set difficulty

Next, we use our NEWS-BALANCED dataset containing 15,454 instances to analyze differences in correction difficulty between confusion sets. As we have already seen in the qualitative analysis above, it is likely that confusion sets like {county,country} are rather difficult, while we might be able to give near perfect feedback for others. Note that when going from unbalanced to the balanced dataset, the false-alarm rate increased an order of magnitude from .001 to .011 indicating that the less common confusion sets are more difficult on average.

In Figure 2, we show false-alarm rate and miss rate for all confusion sets. Results for specific confusion sets may deviate from the average quite a bit. For example, there are confusion sets with no false alarms like {begin, being}, but also {word,

[9] https://www.thecable.ng/does-shiroro-fallen-soldiers-blood-matter/

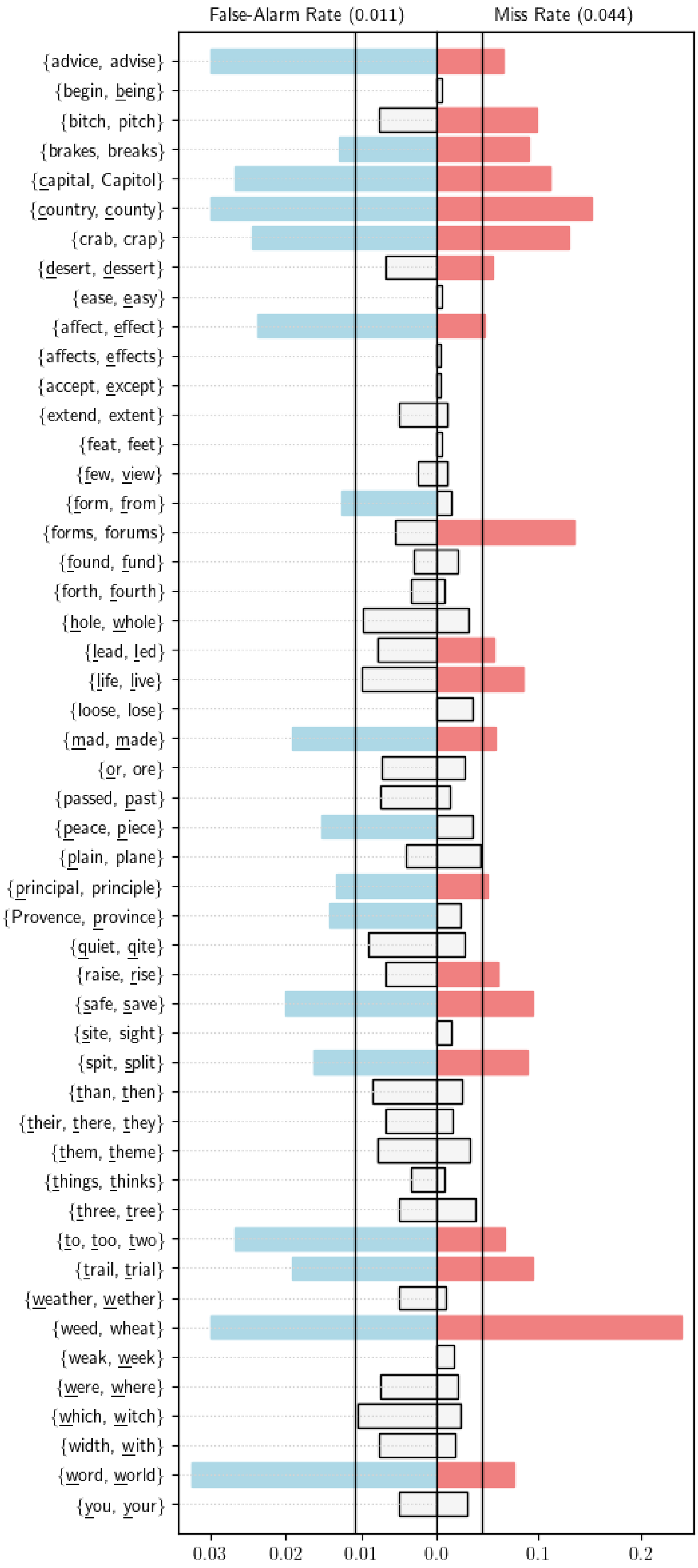


Figure 2: Comparison between false-alarm rate and miss rate on NEWS-BALANCED for all confusion sets. Additional vertical lines show the averages over all confusion sets. Grey bars show values below the average, blue/red bars show above average values.

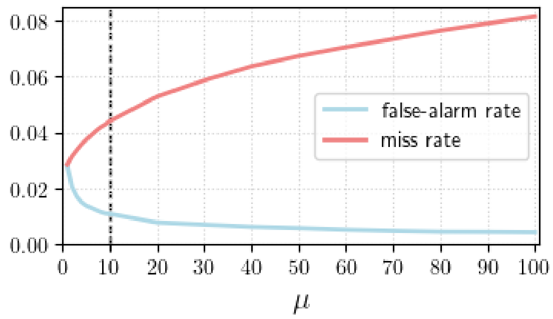


Figure 3: Trade-off between false-alarm rate (lower curve) vs. miss rate (upper curve) based on the threshold parameter μ on the NEWS-BALANCED dataset. The intersections of the plot lines with the vertical, gray line indicate the results of the RWSE detection on the NEWS-BALANCED dataset as presented in Table 2.

world} with a rate three-times the average. We see a similar picture for miss rates: {weed, wheat} stands out with over 20% missed instances.

5.3 Threshold Factor

When designing our detection method, we have somewhat arbitrarily selected a factor of 10 (one order of magnitude) for the μ threshold. Remember that it controls how much more likely a word from the confusion set must be to be considered as a replacement for the original word. We now analyze this choice by computing false-alarm rate and miss rate for different values of μ on all instances of the NEWS-BALANCED dataset. Figure 3 shows the resulting trade-off. Interestingly, our intuitive choice of 10 is already a sensible one striking a good balance between missing out on detection and producing too many false alarms. The chart also shows that e.g. with a value of 100, we could almost entirely eliminate false alarms and only miss about 8% of RWSEs.

5.4 Results on EDUCATIONAL Data

The 18,984 sentences extracted from the CLC-FCE dataset triggered 364 alarms, which we manually annotate. We discard 31 ambiguous instances, where we either would need more context (especially {county,country} cases), the learner language allows for multiple interpretations, or another word outside of the confusion set is more likely. The latter category includes several cases of *"your [yours] sincerely"*, as well as *"Than [Thank] you "* and *"witch [with] NP"*. Here, it is important to remember that our confusion sets are dynamically configurable, which means that a teacher can, when seeing a mistake like this, augment the {which, witch} confusion set into {which, witch, with} depending on whether they consider this confusion to be pedagogically relevant at this point.

Of the remaining 323 instances, 34 are wrong which results in a false-alarm rate of .105. Looking into specific confusion sets, we find that all 14 {quiet,quite} instances are correctly identified which is in line with the rates determined on the NEWS-BALANCED dataset (cf. Figure 2). The same is true for {than,then}, {things,thinks}, {whether,weather}, and {weak,week}. So even if the NEWS dataset underestimate the absolute error rates, it seems to be a good estimate of relative confusion set difficulty. However they are also counterexamples.[10] {their,there,they} is an easy confusion set in the NEWS datasets, but in the EDUCATIONAL instances it is quite hard.

6 Conclusion

In this paper, we tackle the problem of detecting real-word spelling errors in learner text. For that purpose, we present a modern Transformer-based implementation with dynamically configurable confusion sets. We show that our implementation is at least as accurate as earlier approaches when evaluated on news data and when applied on synthetic error data. Our experiments also reveal that learner data is more challenging, but that with our configuration 89% of alarms correctly identify an RWSE. Our analysis also shows that performance varies a lot between confusion sets, but that this could be counter-balanced by adjusting the detection threshold for each confusion set or taking a wider context window into account. We discuss more ideas for future work in the next section together with limitations.

Limitations

The bulk of our experiments is carried out only for English, but as we show by applying it on a small German dataset, the method technically also works for other languages and can be easily adapted.

In our study, we limit the context window to single sentences. Our qualitative analyses have shown that in some cases (few in the NEWS dataset, but quite a few in the EDUCATIONAL data) a wider context would be necessary to resolve the ambiguity. It remains to be empirically tested whether this

[10] This sentence contains a deliberate RWSE. 'they' should be 'there'. Did you spot it while reading?

really would result in fewer such cases.

Another limitation is that we only cover single word errors excluding confusion sets with multiword tokens, like {a life, alive}, or with apostrophes, like {its, it's}. We are also limited by the BERT vocabulary, so that some rarer words are not covered. While this is probably not a major problem in educational texts, as learners are unlikely to produce very infrequent words, multiwords and apostrophes are in the core curriculum. It remains to be investigated on the implementation or configuration level how to treat these cases.

From our analysis, a problem is when a confusion set is triggered, but the correct solution is not in the confusion set. An example was *[Than] you for reading.* We propose to solve this issue, by adding a post-check to our method, where the fill-mask task is run without the confusion set filter, and whenever there is another solution to the masked gap that is overwhelmingly more likely (exact magnitude to be determined empirically), we do not raise an alarm.

Ethics Statement

We do not see any ethical issues with this line of work. Helping people making fewer embarrassing mistakes might have slightly positive effects. In our experiments, we are only using publicly available models and data.

Acknowledgements

We thank Flavio Lötscher for support in compiling the list of confusion sets. We also thank Nico Steinhardt for deploying the demo app. We also thank the reviewers for their insightful comments and constructive feedback.

References

Dimitris Alikaniotis and Vipul Raheja. 2019. The Unreasonable Effectiveness of Transformer Language Models in Grammatical Error Correction. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 127–133, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13–47.

Andrew J. Carlson, Jeffrey Rosen, and Dan Roth. 2001. Scaling up context-sensitive text correction. In Proceedings of the Thirteenth Conference on Innovative Applications of Artificial Intelligence Conference, page 45–50. AAAI Press.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Andrew Golding and Yves Schabes. 1996. Combining Trigram-Based and Feature-Based Methods for Context-Sensitive Spelling Correction. In 34th Annual Meeting of the Association for Computational Linguistics, pages 71–78, Santa Cruz, California, USA. Association for Computational Linguistics.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. Information Processing & Management, 27(5):517–522.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky. 2008. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In Computational Linguistics and Intelligent Text Processing, pages 605–616, Berlin, Heidelberg. Springer Berlin Heidelberg.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Zheng Yuan and Ted Briscoe. 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. Association for Computational Linguistics.

Torsten Zesch. 2012. Measuring Contextual Fitness Using Error Contexts Extracted from the Wikipedia Revision History. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 529–538, Avignon, France. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.