A Benchmark Corpus of English Misspellings and a Minimally-supervised Model for Spelling Correction

Michael Flor
Educational Testing Service
Princeton
NJ 08541, USA
mflor@ets.org

Michael Fried
Queens College
City University of New York
NY 11367, USA
mjf314@gmail.com

Alla Rozovskaya
Queens College
City University of New York
NY 11367, USA
arozovskaya@qc.cuny.edu

Abstract

Spelling correction has attracted a lot of attention in the NLP community. However, models have been usually evaluated on artificially-created or proprietary corpora. A publicly-available corpus of authentic misspellings, annotated in context, is still lacking. To address this, we present and release an annotated data set of 6,121 spelling errors in context, based on a corpus of essays written by English language learners. We also develop a minimally-supervised context-aware approach to spelling correction. It achieves strong results on our data: 88.12% accuracy. This approach can also train with a minimal amount of annotated data (performance reduced by less than 1%). Furthermore, this approach allows easy portability to new domains. We evaluate our model on data from a medical domain and demonstrate that it rivals the performance of a model trained and tuned on in-domain data.

1 Introduction

This paper addresses automatic correction of spelling errors where the misspelled string is not a valid word in the language. Correcting non-word spelling errors has a long history in the natural language processing research (Kukich, 1992). Earlier approaches were evaluated on spelling errors from proprietary corpora of native English texts or artificially generated errors in well-formed texts. While spell checkers today are essential and ubiquitous, dealing with data in a variety of "noisy" domains poses particular challenges to traditional spell checkers. Thus, spelling research has shifted focus primarily to correcting spelling errors in social media data, biomedical texts, and texts written by non-native English writers.

Non-native English speakers account for the majority of people writing in English today, and spelling errors are some of the most frequent error types for these writers (Ng et al., 2014). In some grammatical error correction approaches researchers apply a spell checker prior to running a grammar-oriented correction model (Chollampatt and Ng, 2018; Chollampatt et al., 2016; Rozovskaya and Roth, 2016). In addition to writing-assistance feedback, spelling correction for non-native writers is also utilized in computer-aided language learning applications and in automatic scoring systems (Sukkarieh and Blackmore, 2009; Dikli, 2006; Warschauer and Ware, 2006; Leacock and Chodorow, 2003).

Spelling correction in learner texts is particularly challenging. Non-native writers have higher spelling error rates than native writers (Flor et al., 2015). The types of misspellings produced by these writers typically differ from errors produced by native speakers. While the majority of spelling errors produced by native speakers involve single-character edits (Damerau, 1964), multi-character edits are a lot more common among non-native writers (Flor et al., 2015). Finally, learner data is more likely to contain other errors or non-standard usage in context, which may further complicate error correction (Flor and Futagi, 2012).

Several recent works have specifically addressed spelling correction in learner texts. However, they evaluated either on small data sets (Nagata et al., 2017) or on proprietary corpora (Flor, 2012). Despite several decades of research on spelling, there is still no publicly available large-scale corpus, explicitly and exhaustively annotated for spelling errors. Without such data, it is difficult to compare and track research progress in the field.

This paper makes the following contributions:

- We present a corpus of learner essays, TOEFL-Spell, annotated for spelling errors. This corpus can be used as a benchmark corpus to develop state-of-the-art models for spelling correction (Section 3).

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 76–86
Florence, Italy, August 2, 2019.

- We develop a minimally-supervised approach to spelling correction that combines contextual and non-contextual information (Section 4). We show that inclusion of word embeddings provides information complementary to other contextual features.

- The proposed model is shown to be robust, evaluated on TOEFL-Spell and on an out-of-domain data set of clinical notes. The performance of our model on the clinical data set rivals that of the model trained on a corpus of clinical notes (Section 5).

- Evaluation of the contribution of contextual features shows that contextual information provides an error reduction of about 45%, improving the correction accuracy by 10 points on TOEFL-Spell and by 7 points on the clinical data set.

- Error analysis of the system on TOEFL-Spell and on the clinical data is presented in Section 6.

2 Related Work

A non-word misspelling is a spelling error, such that the produced string is not a valid word in the language. This is different from real-word (context-sensitive) errors, for example confusing "their" and "there" (Wilcox-O'Hearn et al., 2008). This section provides an overview of prior work on correction of non-word spelling errors and availability of corpora for such research.

2.1 Data Sets for Spelling Research

Traditionally, three areas of research have been particularly interested in spelling errors: information retrieval - for misspellings in queries, English language learning - for misspellings made by language learners, and medical information processing - for misspellings in medical documents. Previous work used either proprietary data sets or artificially generated errors. Flor (2012) evaluated on a large corpus of student essays, but the corpus is not publicly available. Toutanova and Moore (2002) and Brill and Moore (2000) similarly evaluated on proprietary data sets of typos collected from native English texts.

Query spelling correction has been an important aspect of research in the domain of information retrieval (Hasan et al., 2015; Chen et al., 2007; Li et al., 2006). The MSR-Bing Web Scale Speller Challenge (Wang and Pedersen, 2011) presented 5500 short queries, with about 10% of them containing typographical errors. Recently, Hagen et al. (2017) presented a large corpus of query misspellings - about 54K queries, with about 9K potential spelling errors. Errors were not explicitly marked; annotators provided alternative formulations, so spelling errors are deduced from comparing the original and revised formulations.

For non-native spelling errors, Nagata et al. (2011, 2017) describe a small corpus (25K words) annotated for various errors, with only 438 spelling error tokens. Mizumoto and Nagata (2017) refer to a newer version of that corpus, with 30K words and 654 spelling errors.

The NUCLE corpus (Dahlmeier et al., 2013) contains 1400 essays written by students at the National University of Singapore, and annotated using twenty seven error codes. In this corpus, spelling errors were included in the *Mechanical errors* category that lumps together quite different types of low-level errors - 'punctuation, capitalization, spelling and typos'. Thus, spelling errors are marked explicitly, but not distinctively.

Heilman et al. (2014) released a corpus of 1511 learner sentences (28K words), judged for grammaticality on an ordinal scale. The JFLEG corpus (Napoles et al., 2017) built on top of that data – for each sentence they added three holistic fluency edits (sentence rewrites) to correct the grammar and also make the original text more fluent. In this corpus, spelling (or other erros) are not explicitly annotated, which makes it difficult to isolate them for spelling correction research. Moreover, the size of this corpus is rather small, and there is no context beyond the sentence level.

The Cambridge Learner Corpus First Certificate in English (FCE) has about 2500 essays (500K words), written by learners taking the English proficiency exam (Yannakoudakis et al., 2011). It was annotated for 80 error types (Nicholls, 2003), including an explicit category for spelling mistakes. However, on closer analysis, one can find that many spelling errors are tagged with other error categories. Thus, its annotation is not directly suitable for spelling correction research.

In the biomedical domain, the largest corpus annotated for spelling errors is a recently released data set of clinical notes (Fivez et al., 2017a), with 873 annotated misspellings in sentence context.

2.2 Approaches to Spelling Correction

Approaches to correcting non-word spelling errors can be broken down into those that only consider the characteristics of the target token when ranking correction candidates, and those that also include the surrounding context. Among the former are those that compute edit distance (Levenshtein, 1966; Damerau, 1964) and phonetic similarity between the misspelling and a candidate correction (Toutanova and Moore, 2002).

A standard approach to correcting non-word spelling errors follows the noisy channel model formulation (Shannon, 1948). It uses edit distance and phonetic similarity between the misspelling and the candidate correction, and the candidate frequency (Kernighan et al., 1990; Church and Gale, 1991; Toutanova and Moore, 2002). Weights for different edit operations are estimated from large training sets of annotated spelling errors. This approach requires a lot of supervision: thousands of annotated errors paired with their corrections are used to estimate probabilities associated with different edits.

The noisy channel model can also incorporate contextual information. For instance, Brill and Moore (2000) ranked candidate corrections by language model scores and reduced the error rate by 73% on correcting artificially-generated errors in the Brown corpus. However, in general, adding new features from a variety of sources is not straightforward in the noisy channel approach.

Contextual features have been used for correcting simulated non-word errors and real-word errors. Carlson and Fette (2007) use a memory-based model with context features estimated from the Google Web1T n-gram corpus (Brants and Franz, 2006). Use of data from the Web for spelling correction was described by Whitelaw et al. (2009) and Chen et al. (2007).

Flor (2012) introduced an approach to ranking candidate corrections that combines edit distance and phonetic distance with contextual cues, and evaluated it on errors made by non-native English speakers. For instance, given *'forst'*, candidate corrections could include *first, forest, frost,* and even *forced*. In a context like *"forst fires in Yellowstone"*, *forest* is a likely candidate. For *"forst in line"*, *first* seems more adequate. That study demonstrated that contextual features significantly improve spelling correction accuracy on an annotated corpus of spelling errors collected from TOEFL and GRE exam essays. It significantly outperformed popular spellers like Aspell and the speller in MS Word (Flor and Futagi, 2012).

3 The TOEFL-Spell Corpus

We base our data set on the publicly available ETS Corpus of Non-Native Written English (Blanchard et al., 2013, 2014), a.k.a. TOEFL11. It consists of essays written for the TOEFL® iBT test, which is used internationally as a measure of academic English proficiency at institutions of higher learning where English is the language of instruction. TOEFL11 contains 12,100 essays from 11 first language backgrounds; 1,100 essays per language, sampled evenly from eight prompts (topics), along with score levels (low/medium/high) for each essay. Each prompt poses a proposition and asks to write an argumentative essay, stating arguments for or against the proposition.

We sampled 883 essays, selecting among those that received medium or high score (low-scored essays are difficult to understand and to annotate). The data set has 296,141 words. Essay length ranges from 168 to 672 words, with an average of 335 words per essay.

The selected essays were annotated by two annotators with linguistic background and prior experience with linguistic annotation. For each essay, an automatic dictionary lookup system highlighted strings that were not found in dictionary. For each highlighted string, the annotator had to determine whether it was indeed misspelled, and to provide an appropriate correction. To ensure the annotation is exhaustive, annotators were also instructed to check for additional misspellings, beyond those highlighted.

The resulting annotation contains **6,121** spelling errors of non-word type, which gives a word error rate of 2.07%. 35 essays had no spelling errors, while the rest had between one and ten errors per essay. The number of unique misspellings is 3,958, and the number of unique correction replacements is 4,016. In most cases, the same error has the same correction; the average number of unique corrections per error is 1.015.

The distribution of misspellings by edit distance to the correct word is presented in Table 1. The majority (82.8%) of errors differ from the correct word by just one character, and an additional 12.6% differ from the correct form by two characters. This is similar to results reported by Flor

Edit distance	Count	Percentage (%)
1	5,066	82.76
2	769	12.56
3	198	3.23
> 3	88	1.45
Total	6,121	100

Table 1: Distribution of errors by edit distance to correct form, in TOEFL-Spell.

et al. (2015) on a different corpus of learner English. Although the majority of errors constitute single-token edits, about 5% (296) are fusion errors (e.g. *'atleast' for 'at least'*).

Randomly chosen, 76 essays were doubly annotated for calculating inter-annotator agreement. A strict criterion was applied for agreement: two annotations had to cover exactly the same segment of text and to specify the same correction. Inter-Annotator Agreement was 95.6%. (Note that Kappa statistic cannot be applied to error correction, as there are too many different responses).

The full set of annotations for TOEFL-Spell is released and made available for research.[1]

4 The Spelling Correction Model

In this section, we present our benchmark model of spelling correction, which extends the model of Flor (2012). The spelling correction task consists of three subtasks: detection, generating candidate corrections, and ranking of the candidates.

4.1 Error Detection

Detection of non-word misspellings is performed using a dictionary (lexicon). Tokens that are not in the lexicon are considered to be misspelled. We use a dictionary that consists of 140,000 single words (including inflections), 100,000 multi-word terms, and 130,000 names (including names and surnames from various countries). The dictionary includes both American and British spelling variants, common acronyms, and foreign words. The dictionary includes lexica from WordNet,[2] the SCOWL project,[3] names from US Census Data,[4], Wikipedia lists[5], and various sources on the Web.

[1] https://github.com/EducationalTestingService/toefl-spell
[2] https://wordnet.princeton.edu/
[3] http://wordlist.aspell.net/dicts/
[4] 2010 Surnames, on census.gov
[5] https://en.wikipedia.org/wiki/Category:Names_by_language

Feature name	Description
Non-contextual features	
Orthographic similarity	Inverse edit distance
Phonetic similarity	Inverse edit distance of phonetic representations
Word frequency	Candidate word frequency in language
Contextual features	
N-gram support	N-gram counts in a 4-word window (from corpus)
Dejavu	Is the candidate found elsewhere in same essay
DejavuSM	Is the candidate found as candidate for other errors in same essay
Word embeddings	Using word embeddings to estimate candidate word's relatedness to context

Table 2: Description of all the features used in the candidate ranking module.

4.2 Candidate Generation

Candidates are generated using the dictionary described above. Candidates include all dictionary words within edit distance that does not exceed half of the length of the misspelled string, with a maximum distance of 6 characters. Both single-token and multi-token candidates are generated, to allow for correction of fusion errors. For each misspelled token, hundreds of correction candidates are generated, using the Ternary Search Tree data structure (Bentley and Sedgewick, 1997).

4.3 Ranking of Candidate Corrections

The ranking step is the most challenging one and is the focus of the most work on non-word spelling correction (Fivez et al., 2017b). Our model uses both the features of the misspelling+candidate pair and the contextual information. The former include orthographic similarity, phonetic similarity, and candidate word frequency. The contextual information includes n-gram support, an estimate of potential re-use of words in text, and word embeddings. The features are listed in Table 2.

Orthographic similarity is computed as inverse edit distance, $1/(eDist + 1)$, where $eDist$ is the edit distance (including transpositions) between the misspelling and the correction candidate (Levenshtein, 1966; Damerau, 1964).

Phonetic similarity reflects the intuition that a good correction should be phonetically similar to the misspelling. It is computed as $1/(eDistPh + 1)$, where $eDistPh$ is the edit distance between the phonetic representation of the misspelling and the phonetic representation of the candidate. Phonetic representations are computed using the Double-Metaphone algorithm (Philips, 2000).

Candidate frequency. A more frequent word is more likely to be the intended word than a rare word (Kernighan et al., 1990). Unigram word frequency is computed for each candidate using the English Wikipedia corpus.

N-gram support. For each correction candidate, all n-grams in the window of four context words on each side are taken into account by the n-gram support feature. We use co-occurrence counts computed from the English Wikipedia corpus and weighted as the Positive Normalized PMI scores (PNPMI). Normalized PMI was introduced by Bouma (2009), we adapt it as:

$$log_2 \frac{p(c, ngram)}{p(c)p(ngram)} \Big/ (-log_2 p(c, ngram)) \quad (1)$$

PNPMI maps all negative values to zero. For each candidate c, all n-grams of lengths 2-to-4 words in the context window are generated, and the PNPMI values of each $c, ngram$ pair are added.

Dejavu. This feature considers essay-wide context and rewards a candidate that appears in the same essay. Each occurrence of the candidate (or its inflection) in the text strengthens the candidate by the amount $1/sqrt(1 + distance)$, where distance is the number of tokens between the misspelling and the position of the candidate in text.

DejavuSM is a feature that caters for systematic misspellings, when a word is misspelled throughout the essay (Flor, 2012). For each candidate correction, we search in the lists of candidate corrections of other misspelled tokens in the text. Each time the candidate or its inflection is found in another list, the candidate is strengthened with a score of $S_{CC}/sqrt(1 + distance)$, where S_{CC} is the current rescaled overall strength of the corresponding candidate in the other list.

Word embeddings have shown a lot of success in many NLP applications, especially for estimation of semantic relatedness (Levy and Goldberg, 2014). We use word embeddings to score the contextual fit of correction candidates in the local context of a misspelling. The idea is that for a misspelling like *"roat"*, a correction to *"road"* should be strengthened if a word like *"drive"* is found in the vicinity. Given a misspelled token, we define a window of ±15 tokens around it. For every candidate, we compute the cosine similarity between the embedding vector of the candidate and the vector of each context word, and sum those values. This is the vector-based contextual fit score for the candidate. We use the *word2vec* vectors with 300 dimensions, pre-trained on 100 billion words of Google News (Mikolov et al., 2013).[6]

Ranking of candidates. For each misspelled token, the feature scores of its candidate corrections are *normalized*, by dividing the score of the candidate feature by the highest-scoring candidate on that given feature. The final score for each candidate correction is computed as a weighted sum of the feature scores for the candidate:

$$CandidateScore = \sum_f w_f \cdot S_f$$

where f ranges over the seven feature types used, S_f is the normalized score of the current candidate by feature f, and w_f is the predefined weight of the feature. Learning of weights is described in Section 5.

Our *baseline* system implements all the features, with the exception of word embeddings. Due to the feature formulation, each feature group (e.g. orthographic similarity) requires only one weight. Feature weights for the baseline model are adopted from Flor (2012), where they were manually tuned. In the present work, feature weights are automatically learned with a linear machine learning algorithm. We use two linear classifiers – Logistic Regression and Averaged Perceptron.

5 Experiments

We address the following research questions:

- How does the model compare to a *baseline* system?
- What is the contribution of individual features, especially those that provide contextual information?
- How much training data is needed to learn a robust model?
- How does the model behave on out-of-domain data?

[6] https://code.google.com/archive/p/word2vec

5.1 Experiments on TOEFL-Spell

First, we present results on error detection. The system detected all 6,121 misspellings and flagged 43 additional words (false positives). Thus, the detection recall is 100%, precision is 99.3% and F1 score is 99.65%. This result applies to all experiments with the TOEFL-Spell data set. The candidate generation performance is over 99%, i.e. for over 99% of the errors a valid correction is generated in the list of candidates. Note that in the candidate generation stage, an average of 213 candidate corrections is generated for each misspelling in the TOEFL-Spell corpus.

We now evaluate the performance of the candidate ranking component, checking whether the top-ranked candidate is indeed the gold correction. The *baseline* system implements all the features, except word embeddings, and uses weights from Flor (2012). For the new approach we add the feature computed with word-embeddings. Feature weights are learned automatically, using linear classifiers – Logistic Regression and Averaged Perceptron.

We address the first research question above, using the TOEFL-Spell corpus in a five-fold cross-validation. Results are presented in Table 3. Each of the classifiers outperforms the baseline, and the differences are statistically significant (by two-proportions z-Test). The difference between Perceptron and Logistic Regression is not significant. The Perceptron algorithm is the best model, with over 2 points of absolute improvement, which is an error reduction of 15%.

Model	Accuracy
Baseline (Flor, 2012)	85.97
Logistic Regression (this work)	87.83
Perceptron (this work)	88.12

Table 3: Error correction results for the baseline model and two linear classifiers on the TOEFL-Spell data set. Classifiers outperform the baseline (p<0.002).

Contribution of contextual and non-contextual features. To assess the contribution of individual information sources, we perform feature ablation, by removing one feature at a time. Results are presented in Table 4. The top part of the table shows feature ablation for non-contextual features. The most useful is the orthographic similarity: its removal results in a drop of almost 10 points. Among the contextual features, n-gram support and word2vec prove to be the most useful. Notably, n-gram features and word2vec supply complementary information, and removing each one of those results in a drop in performance. Interestingly, the dejavu and dejavuSM features provide almost no improvement; this result contradicts the finding by Flor (2012). Eliminating all contextual features lowers the performance by more than 10 points, to 77.93%. This demonstrates that contextual features have a substantial contribution. Overall, about 45% of the inadequate corrections produced by the non-contextual model can be corrected by adding context information.

Feature set	Accuracy
Without *orthographic sim.*	79.84*
Without *phonetic sim.*	86.47*
Without *word freq.*	88.07
Without *dejavu*	88.07
Without *dejavuSM*	88.01
Without *word2vec*	86.65*
Without *ngram support*	82.62*
Without contextual features	77.93*
Without non-contextual features	65.63*
All features	**88.12**

Table 4: Feature ablation performance (error correction accuracy %) on TOEFL-Spell. All models are trained with the Perceptron algorithm in 5-fold cross-validation. Values marked by * differ significantly from the value for *All features*, with $p < 0.003$.

How much training data is needed for a robust model. We train the Perceptron classifier, varying the amounts of training data between 5% and 75% of the entire data set. We similarly perform experiments using 5-fold cross-validation, with the exception that we use less data for training each time. 5% of the training data corresponds to about 240 spelling errors in training. Table 5 demonstrates that even with the smallest training set the

Amount of training data	Accuracy
5%	87.67
10%	87.73
20%	87.86
50%	88.04
75%	88.07
100%	88.12

Table 5: Error correction performance (accuracy %) of the Perceptron classifier trained on different amounts of data, on TOEFL-Spell in 5-fold cross-validation.

drop in performance is less than 1%. In fact, the differences between the models are not significant.

We emphasize that the noisy-channel model requires thousands of examples to estimate the weights of individual edits. In this paper, orthographic similarity is represented as a single feature; thus only one weight is estimated (as opposed to about 1000 weights for character pairs). The same is done for our other features, which allows us to train with a small amount of supervision, couple of hundred of errors.

5.2 Out-of-domain Evaluation

We evaluate the model on a data set from a very different content domain – clinical medical records. The genre of clinical free text poses an interesting challenge to the spelling correction task, since it is notoriously noisy (Fivez et al., 2017a; Lai et al., 2015).

Clinical corpora typically contain higher spelling error rates of 7% to 10%, while in native English text error rates usually range between 0.1% and 0.4% (Ruch et al., 2003). Clinical text contains domain-specific terminology and language conventions. Clinical data, in addition to highly domain-specific vocabulary, can also be characterized by a large amount of noise, e.g. the use of non-standard phrases and abbreviations and is thus particularly challenging (Fivez et al., 2017a). These properties can render traditional spell checkers less effective (Patrick et al., 2010).

We use a data set of clinical notes extracted from the large MIMIC-III medical corpus (Johnson et al., 2016). The data set contains 873 manually annotated misspellings (Fivez et al., 2017a). The distribution of errors in this data set in terms of the edit distance is very similar to that in TOEFL-Spell (see Table 1). In particular, 83% of errors have edit distance of 1 to the correction, while another 15% have an edit distance of 2.

The state-of-the-art results on this data set are reported by Fivez et al. (2017a). Their model is tuned on artificially generated spelling errors and trained on word and character embeddings from MIMIC-III (note that MIMIC-III is the superset of the annotated clinical data set). Their model outperforms off-the-shelf spelling correction tools (Aspell) and the noisy channel model. Similarly to (Fivez et al., 2017a), we accommodate to the medical domain by enhancing the dictionary with a comprehensive medical lexicon (the UMLS® SPECIALIST Lexicon.[7])

Model	**Accuracy** *off-the-shelf*	**Accuracy** *completed*
Fivez et al. (2017a)	88.21	93.02
Logistic Regression	87.40	89.35
Perceptron	87.63	89.00

Table 6: Clinical corpus: Performance (accuracy %) of the state-of-the-art system that uses in-domain data, and of the models proposed in this work.

Features	**Accuracy**
Without *orthographic sim.*	58.88
Without *phonetic sim.*	85.68
Without *word freq.*	87.51
Without *dejavu*	87.06
Without *dejavuSM*	87.74
Without *word2vec*	84.88
Without *ngram support*	85.22
Without contextual feats	80.18
Without non-contextual feats	31.73
All features	**87.63**

Table 7: Feature ablation performance (accuracy %) on the clinical data set. All models are trained with the Perceptron algorithm on TOEFL-Spell data.

Fivez et al. (2017a) note that some of the required rare corrections were not available even in the medical lexicon. For this reason, they report two versions of results: *off-the-shelf* (using general+medical dictionaries), and *completed lexicon* (where additional rare terms from the annotations were added to the dictionary).

Results for off-the-shelf evaluation are reported in Table 6. Our models were trained on TOEFL-Spell (the same models reported in Table 3). Note that our n-gram and embedding features are also not from the clinical domain. In the off-the-shelf evaluation, our models achieve performance that is comparable to the state-of-the-art system that used in-domain data and was tuned on the clinical corpus. In the completed lexicon evaluation, the Fivez et al. system is better: it obtained a score of 93.02 vs. 89.35 for our Perceptron algorithm. We believe that the off-the-shelf performance reflects a more realistic scenario, as manually adding candidates to the dictionary introduces bias. We further discuss this in the next section.

[7] https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html

Finally, we evaluate the contribution of each information source on the clinical data (Table 7). Orthographic similarity is the most useful feature, just as it is in the TOEFL-Spell data set, and removing it results in a very big performance drop (almost 30 points). Unsurprisingly, the orthographic similarity feature works well cross-domain. The least helpful features are word frequency, dejavu, and dejavuSM. This is consistent across the two data sets. The word2vec feature provides a slightly better improvement on the clinical data (3 points vs. 2 on TOEFL-Spell), while the n-gram feature performs slightly worse (only 2 points improvement, compared to 6 on TOEFL-Spell). Overall, contextual features contribute 7 points here versus 10 on TOEFL-Spell. This result is expected given that contextual features are estimated on out-of-domain data.

In sum, the experiments on the clinical data set demonstrate that our model is robust and competitive on out-of-domain data. This also stresses the value of the TOEFL-Spell data set, on which our model was trained.

6 Error Analysis

We perform error analysis on both data sets. We first consider cases where the gold correction was not selected as the top candidate. For the TOEFL-Spell data set, our best system places the gold correction at the top of the ranked list in 88% of the cases. If we consider the top five candidates, the system finds the gold correction in 96.7% of the cases. We investigate the cases where the top candidate is different from the gold. In 15.25% of the cases, the top candidate and the gold are inflectional variants of the same lemma (e.g. error: *updations*, gold: *updates*, system-best: *updating*). In 11.4% of cases, the top candidate and the gold are derivationally related (e.g. error: *elastico*, gold: *elasticity*, system-best: *elastic*). In 4% of cases, the top candidate is a close variant of the gold (e.g. error: *donot*, gold: *do not*, system-best: *don't*), or a US/UK spelling variant (e.g. error: *bahaviours*, gold: *behaviours*, system-best: *behaviors*).

For the clinical data set, the system's top suggestion is correct in 87.6% of the cases. The gold correction appears among the top five candidates in 96.7% of the cases (with off-the-shelf dictionaries). In 29.6% of the cases with an incorrect top candidate, the top candidate and the gold correction are inflectional variants of the same lemma, in 14.4% of the cases they are derivationally related, and in 3% of the cases, the top candidate simply has an alternative spelling (e.g. *cyclosporin* and *ciclosporin*). Overall, in 43% of the cases the system selects a morphological variant of the gold correction. This number is lower for the TOEFL-Spell corpus (25%).

We also checked why, in the completed lexicon evaluation on clinical data, our model does not perform as well as the one by Fivez et al. (2017a). It turns out that our model has poor accuracy on the specially added words (41.38%). Further inspection shows that these manually added words are extremely rare medical terms. As a result, contextual features do not fire on them. We expect that adding medical corpora to train word embeddings will solve this issue.

Finally, we provide some examples of errors that our system managed to correct with contextual information but failed to correct without context. An example from the clinical data set: *"was thought to be cold agglutin hemolytic anemia..."*. Without context, the system chooses *agglutin* → *gluten*. With context, the system chooses *agglutin* → *agglutinin*, because *"cold agglutinin"* happens to be a strong collocation. An example from the TOEFL-Spell data set: *"countries such as england, fance and the usa are..."*. Without context, the system prefers *fance* → *fence*, but with context, it correctly chooses *fance* → *france*.

7 Conclusions

This paper addressed the problem of correcting non-word spelling errors, with a focus on errors occurring in noisy natural data. We presented TOEFL-Spell, a publicly-available large data set of authentic misspellings annotated in context. This data set should facilitate further research on spelling correction for noisy data.

We also presented a minimally-supervised model for spelling correction that utilizes non-contextual and contextual features, and does not require a lot of training data. The model demonstrated a state-of-the-art performance on data sets from two noisy domains: learner data and clinical notes. On the latter, competitive performance was achieved, compared to a model developed specifically for the medical domain and trained on in-domain clinical data. We plan to extend this model for handling real-word spelling errors.

Acknowledgments

We thank the anonymous reviewers for valuable comments that helped us to improve this paper.

References

Jon L. Bentley and Robert Sedgewick. 1997. Fast algorithms for sorting and searching strings. In *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA97*, pages 360–399. ACM.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English. Research Report ETS RR13-24*. Educational Testing Service, Princeton, NJ, USA.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. *ETS Corpus of Non-Native Written English, Catalog No. LDC2014T06*. Linguistic Data Consortium, Philadelphia, PA, USA.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference*, pages 31–40, Tbingen. Gunter Narr Verlag.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293.

Andrew Carlson and Ian Fette. 2007. Memory-based context-sensitive spelling correction at web scale. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Qing Chen, Mu Li, and Ming Zhou. 2007. Improving query spelling correction using web search results. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 181–189, Prague, Czech Republic. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

Shamil Chollampatt, Kaveh Taghipour, and Hwee Tou Ng. 2016. Neural network translation models for grammatical error correction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16*, pages 2768–2774, New York, NY, USA. AAAI Press.

Kenneth W. Church and William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing*, 1:93–103.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Frederick Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):659–664.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5:4–35.

Pieter Fivez, Simon Suster, and Walter Daelemans. 2017a. Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings. In *BioNLP 2017*, pages 143–148, Vancouver, Canada,. Association for Computational Linguistics.

Pieter Fivez, Simon Łuster, and Walter Daelemans. 2017b. Unsupervised Context-Sensitive Spelling Correction of English and Dutch Clinical Free-Text with Word and Character N-Gram Embeddings. In *Arxiv*.

Michael Flor. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL)*, 53(3):61–99.

Michael Flor and Yoko Futagi. 2012. On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 105–115, Montréal, Canada. Association for Computational Linguistics.

Michael Flor, Yoko Futagi, Melissa Lopez, and Matthew Mulholland. 2015. Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus. In *Learner Corpus Research: LCR2013 Conference Proceedings*, volume 6 of *Bergen Language and Linguistic Studies (BeLLS)*, pages 107–132.

Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, and Benno Stein. 2017. A Large-Scale Query Spelling Correction Corpus. In *40th International ACM Conference on Research and Development in Information Retrieval (SIGIR 17)*, pages 1261–1264. ACM.

Saša Hasan, Carmen Heger, and Saab Mansour. 2015. Spelling correction of user search queries through statistical machine translation. In *Proceedings of the*

2015 Conference on Empirical Methods in Natural Language Processing, pages 451–460, Lisbon, Portugal. Association for Computational Linguistics.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035+.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Papers presented to the 13th International Conference on Computational Linguistics (COLING 1990)*, volume 2, pages 205–210.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24:377–439.

Kenneth H. Lai, Maxim Topaz, Foster R. Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 15:188–195.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and Humanities*, 37:389–405.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Mu Li, Muhua Zhu, Yang Zhang, and Ming Zhou. 2006. Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1025–1032, Sydney, Australia. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomoya Mizumoto and Ryo Nagata. 2017. Analyzing the Impact of Spelling Errors on POS-Tagging and Chunking in Learner English. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 54–58, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ryo Nagata, Hiroya Takamura, and Graham Neubig. 2017. Adaptive Spelling Error Correction Models for Learner English. *Procedia Computer Science*, 112:474–483.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219, Portland, Oregon, USA. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 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. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581.

Jon Patrick, Mojtaba Sabbagh, Suvir Jain, and Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 2–8.

Lawrence Philips. 2000. The double-metaphone search algorithm. *Dr. Dobb's Journal*.

Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2205–2215, Berlin, Germany. Association for Computational Linguistics.

Patrick Ruch, Robert Baud, and Antoine Geissbuhler. 2003. Using lexical disambiguation and namedentity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, 29:169–184.

Claude Shannon. 1948. A mathematical theory of communications. *Bell Systems Technical Journal*, 27:623–656.

Jana Sukkarieh and John Blackmore. 2009. C-rater: Automatic content scoring for short constructed responses. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*, pages 290–295.

Kristina Toutanova and Robert Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kuansan Wang and Jan Pedersen. 2011. Review of MSR-Bing web scale speller challenge. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR 2011)*, pages 1339–1340.

Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2):157–180.

Casey Whitelaw, Ben Hutchinson, Grace Y Chung, and Ged Ellis. 2009. Using the Web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 890–899, Singapore. 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 *Proceedings of CICLing-2008*, pages 605–616.

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.