Do Word Embeddings Capture Spelling Variation?

Dong Nguyen[♠,♡], **Jack Grieve**[♣],
[♠]The Alan Turing Institute, London, UK
[♡]Department of Information and Computing Sciences, Utrecht University, the Netherlands
[♣]Department of English Language and Linguistics, University of Birmingham, UK
d.p.nguyen@uu.nl, J.Grieve@bham.ac.uk

Abstract

Analyses of word embeddings have primarily focused on semantic and syntactic properties. However, word embeddings have the potential to encode other properties as well. In this paper, we propose a new perspective on the analysis of word embeddings by focusing on spelling variation. In social media, spelling variation is abundant and often socially meaningful. Here, we analyze word embeddings trained on Twitter and Reddit data. We present three analyses using pairs of word forms covering seven types of spelling variation in English. Taken together, our results show that word embeddings encode spelling variation patterns of various types to some extent, even embeddings trained using the skipgram model which does not take spelling into account. Our results also suggest a link between the intentionality of the variation and the distance of the non-conventional spellings to their conventional spellings.

1 Introduction

Word embeddings play a key role in many NLP systems. Unfortunately, embeddings are opaque: It is difficult to interpret the individual vector dimensions and to understand which factors contribute to the relational similarity between words nearby in space. There has been an increasing interest in understanding what word embeddings are encoding, for example, to understand their impact on the performance of a wide array of NLP systems (Rogers et al., 2018). Moreover, this is especially important when embeddings are used as research objects themselves, e.g., to study biases in society over time (Garg et al., 2018). The body of work on analyzing embeddings has focused primarily on semantic and syntactic properties (Baroni et al., 2014; Gladkova et al., 2016; Levy and Goldberg, 2014; Mikolov et al., 2013b). Instead, we propose a new perspective on word embeddings by asking how *spelling variation* is encoded.

In social media, linguistic variation is abundant (Eisenstein, 2013; Tatman, 2015). For example, non-conventional spellings for *nothing* are *nothin*, *nuthin*, *nooooothing*, *nithing*, and so on. Much work in NLP has focused on normalizing spelling variation (Han and Baldwin, 2011; Liu et al., 2011). Similarly, recent work by Piktus et al. (2019) aims to push embeddings of misspelled words closer to the embeddings of their conventional forms. However, these approaches can remove valuable social (Eisenstein, 2013) and semantic (Grieve et al., 2017) signals. Crucially, many non-conventional spellings are not misspellings: by deliberately deviating from conventional spelling norms, writers create social meaning (Sebba, 2007). For example, a certain spelling may be used to evoke intimacy or to index a certain region. Capturing spelling variation patterns is therefore important for many applications, e.g., when analyzing social phenomena (Nguyen et al., 2016).

To illustrate, Figure 1 shows skipgram embeddings of *nothing* and spelling variants projected on a 2D plane using t-SNE (van der Maaten and Hinton, 2008). This figure clearly suggests structure in the embedding space as forms with different types of spelling variation are pulled apart. On the middle right side we find lengthened forms (e.g., *nooooooothing*), which are often used for emphasis, and at the top we find forms that appear to be misspellings (e.g., *nithing*) and the conventional form (*nothing*).

Proceedings of the 28th International Conference on Computational Linguistics, pages 870–881
Barcelona, Spain (Online), December 8-13, 2020

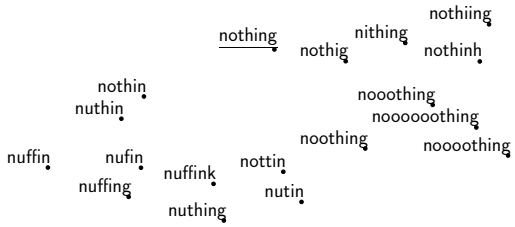


Figure 1: *nothing* and spelling variants in Reddit (skipgram embeddings)

In the lower left corner, we find instances of g-dropping (e.g. *nothin*) and forms reflecting dialect pronunciation (e.g., *nuffin*, which presumably reflects a dialect pronunciation of the voiceless interdental fricative */θ/*). Because spelling is not taken into account by skipgram, this suggests that these forms are used in different contexts. Indeed, research has associated a range of contextual factors with the use of specific types of spelling variations (Eisenstein, 2015). We therefore go beyond looking at whether non-conventional forms are close to their conventional form in the embedding space. Instead, we ask whether embeddings capture fine-grained patterns of spelling variation, such as information about the type of spelling variation (e.g., whether the final 'g' is dropped, or whether the vowels are omitted).

Why does spelling variation matter? If embeddings are supposed to represent language use, we should also expect them to represent meaning associated with the choice of a specific spelling. Non-conventional spellings often carry social meaning, e.g. reflecting and constructing social identities. But, crucially, this has been neglected in NLP at large and in representation learning more specifically, ignoring the fact that in language *it doesn't only matter what is said, but also how it's said.* If the relationship between embeddings of conventional and non-conventional forms is meaningful, then this opens up a range of opportunities of using embeddings for exploring questions in computational social science and sociolinguistics. Moreover, there are many downstream applications that would benefit (e.g., community detection, modeling conversation dynamics). And finally, NLP needs to move away from focusing almost exclusively on so-called "standard language", resulting in systems that do not work well for many social groups. Treating spelling variation as a meaningful phenomenon is one step in this direction.

Contributions Our contributions are (i) a new perspective on the analysis of word embeddings by focusing on spelling variation, (ii) a new dataset with seven common types of spelling variation to analyze and evaluate word embeddings (Section 3), and (iii) an empirical investigation using three analyses, revealing that spelling variation is indeed encoded to some extent—even with the skipgram model, which does not take spelling into account—and that there are differences between the types of spelling variation (Section 4). The code is available on https://github.com/dongpng/coling2020.

2 Related Work

NLP research on spelling variation has mostly focused on text normalization (Han and Baldwin, 2011; Liu et al., 2011) and automatically extracting lexical variants (Gouws et al., 2011). Studies within computational sociolinguistics have analyzed the patterns and functions of spelling variation, e.g., in Twitter (Tatman, 2015), and research has suggested a deep connection between phonological and spelling variation (Eisenstein, 2015). Furthermore, Thurlow and Brown (2003) observed that spelling variation also personalizes and informalizes SMS messages. We look at spelling variation but focus on how spelling variation patterns are encoded in embeddings. There has been much interest in analyzing neural representations and more broadly neural networks (Belinkov and Glass, 2019). Besides word embeddings, representations for higher-level units such as sentences have also been analyzed (Conneau et al., 2018). Our study differs from work in this space not only because it is based on spelling but also because it is based on comparing forms with the same referential meaning. In contrast, many evaluation studies (e.g., Gladkova et al. (2016)) use pairs that are not semantically (*cat:kitten*) or grammatically (*argue:argument*) equivalent. Our study is more similar to Niu and Carpuat (2017), who used paraphrase pairs to investigate formality information in embeddings, and Shoemark et al. (2018), who focused on three language variety pairs. We are not aware of studies focusing on *spelling*.

Type	# Pairs	Cohen's κ	Agreement (%)	Examples
Lengthening	258	0.634	94%	*andddddd/and*
Swapped characters	85	0.865	94%	*myslef/myself*
Common misspellings	76	*NA, based on existing list*		*existance/existence*
G-dropping	233	0.672	88%	*goin/going*
Vowel omission	45	0.898	98%	*grl/girl*
Nearby character substitution	40	0.668	92%	*intetesting/interesting*
British vs. American spelling	57	0.875	96%	*flavour/flavor*

Table 1: Types of spelling variation

3 Datasets

This section describes the datasets and the types of spelling variation we use in our analyses.

3.1 Data Collection

Spelling variation is especially common in social media. We therefore use data from two popular social media platforms. Our datasets complement each other in terms of their characteristics (Twitter versus Reddit, location constrained versus global English). We focus on posts in English.

Twitter One year of geotagged tweets constrained to the London area (May 2018–April 2019). The tweets were tokenized using Twokenizer (O'Connor et al., 2010). The dataset contains 14.1M tweets and a vocabulary of 100.2k words (min. freq. count: 20).

Reddit Six months of Reddit comments (May–Oct 2018), posted in the top 200 subreddits based on comment counts (after manually excluding a few country specific subreddits to minimize non-English content). The dataset contains 269M posts and a vocabulary of 576k words (min. freq. count: 20).

3.2 Types of Spelling Variation

Our analyses are based on pairs of forms that generally have the same referential meaning but different spellings. In this paper, we refer to the form *without* the spelling variation as the **conventional form** (e.g., *nice*) and the form *with* the spelling variation as the **non-conventional form** (e.g., *niiiiice*).[1] We focus on seven common types of spelling variation (see also Table 1), based on observations in our data as well as based on types identified in related work:

- **Lengthening**: The repetition of characters, e.g., *thaaaaanks* instead of *thanks*. Lengthening is often used for emphasis and it is a useful signal for sentiment detection (Brody and Diakopoulos, 2011). It has also found to be used more by younger Twitter users (Nguyen et al., 2013). We automatically searched for forms with their final character repeated (sequences of the same character of length ≥ 3) and for forms where the lengthening occurred by repeating an internal vowel.
- **Swapped characters**: Pairs for which the difference between the two forms is the swapping of two characters (Pennell and Liu, 2011). One of the forms is required to be included in an English wordlist from Aspell[2] (instances of metathesis[3]).
- **Common misspellings**: We use a list of common misspellings.[4]
- **G-dropping (-ing vs. -in):** G-dropping is a phonological alternation that is common in all forms of English. It has a long history of research in sociolinguistics, showing variation among many social factors, including social class and formality (Campbell-Kibler, 2006; Levon and Fox, 2014; Tamminga, 2017). We automatically searched for pairs of *-ing* and *-in* and manually selected the ones that were genuine cases of g-dropping, e.g., excluding instances like *turin* and *turing*.

[1] These are sometimes also referred to as "standard" and "non-standard" forms. We prefer "conventional" and "non-conventional" forms, as some of these forms are very common but not included in dictionaries.

[2] http://aspell.net/

[3] Often not deliberate in writing, but instances of metathesis sometimes become standard over time.

[4] https://en.oxforddictionaries.com/spelling/common-misspellings, accessed 24 Oct 2018.

- **Vowel omission**: For example, *thnks* versus *thanks* (Shortis, 2016; van der Goot et al., 2018).
- **Nearby character substitution**: Pairs for which one form is created by replacing a character by another character at an adjacent key (assuming a QWERTY keyboard layout).
- **British vs. American spelling**: Pairs based on heuristics from online sources.[5] We search for *-our* vs. *-or*, *-yse* vs. *yze*, doubling of l (*-elled* vs. *eled*, etc.), *-ogue* vs. *og* and *-ence* vs. *ense*.[6]

We note that there are many other types of spelling variation not considered here, such as number replacement (e.g., *4ever* vs. *forever*) and phonetic respellings (e.g., *nite* instead of *night*) (Shortis, 2016; Tagg, 2009). For each type, we selected pairs consisting of a non-conventional (e.g., *backkkk*) and a conventional form (*back*). We only included pairs with both forms occurring at least 20 times in each dataset, to have the same pairs for both datasets. Within each of our seven types of spelling variation, a conventional form occurs only once.

Pairs were identified automatically using heuristics, but the final selection was based on manual filtering. Two expert annotators annotated 50 pairs of each type,[7] resulting in substantial to near perfect agreement, see Table 1. For each pair, example posts were available to help interpretation. Disagreement occurred with cases where it was unclear whether both forms had the same referential meaning, for example with *teeets*/*tweets* (substitution for a nearby character) and *plantin*/*planting* (g-dropping, where one annotator was concerned that *plantin* was also a misspelling of *plantain*). Disagreement also occurred when the conventional form was unclear, e.g., for *thiccc* (lengthening). One annotator then went through all automatically identified pairs to select valid pairs. The final selection was then checked by a second annotator. We aimed for precision and therefore excluded ambiguous pairs.

3.3 Extra-Linguistic Variation

We analyze whether the types of spelling variation exhibit extra-linguistic variation in our data, by looking at estimated age distributions of Twitter users and subreddit distributions in Reddit.

Twitter For each form in our lists with spelling variations, we randomly sample five Twitter users. We estimate the demographics of these users using the M3 model (Wang et al., 2019), based on information such as the profile image and user name. The model notably does not make use of the language in tweets. Tweets with lengthened or g-dropped forms were more often written by younger Twitter users (Table 2). For example, 17% of the sampled users who used lengthened forms were estimated to be ≤ 18 years, as opposed to only 10% of the sampled users who used the corresponding conventional forms.

When we look at British versus American spellings, the trend is reversed. Our Twitter dataset was constrained to the London area. Of the users who used the British spelling, a larger proportion was estimated to be older (e.g. 56% of the users was estimated to be $>= 40$ years). On the other hand, of the users who used an American spelling, a larger proportion was estimated to be younger.

Type	**≤18**		**19–29**		**30–39**		**≥40**	
	alt	conv	alt	conv	alt	conv	alt	conv
lengthening	17%	10%	50%	27%	17%	25%	16%	38%
g-dropping	16%	7%	37%	27%	21%	27%	26%	40%
British. vs. American	5%	10%	18%	23%	22%	23%	56%	44%

Table 2: Distributions of spelling variants across estimated age distributions in Twitter, for alternative (alt) and conventional (conv) forms. For example, the values under the "alt" columns indicate the proportion of users (out of all sampled users who used an alternative form), who were estimated to fall in the corresponding age category.

[5] https://www.oxfordinternationalenglish.com/differences-in-british-and-american-spelling/ and https://en.wikipedia.org/wiki/American_and_British_English_spelling_differences

[6] Note that this type of variation is different from the others as both British and American spelling forms are conventional in their geographic regions. In our analyses, we use the forms with American spelling as the 'conventional' forms.

[7] Except for the common misspellings list, as this came from an existing resource.

Reddit For each spelling variation type, we rank the subreddits based on their ratio of relative frequencies of conventional vs. non-conventional forms, see Table 3. Unsurprisingly, British forms occur most in UK-focused subreddits. For example, in the *ukpolitics* subreddit 94.88% of all occurrences of the words in the British/American spelling list are written with a British spelling. Overall, we also find that non-conventional forms occur with low frequency in *TranscribersOfReddit*, a subreddit for curating and supporting transcriptions of Reddit content, e.g., to assist users who rely on text to speech. Moreover, in this subreddit many posts were written by bots. In sum, we find that certain types of spelling variation indeed occur more frequently in certain contexts in our data. As a result, some of this patterning may also be captured in the embeddings.

Type	Top subreddits	Bottom subreddits
Lengthening	thanosdidnothingwrong (0.04%) teenagersnew (0.04%)	legaladvice (0.00%) changemyview (0.00%)
Swapped characters	Bitcoin (0.51%) CryptoCurrency (0.38%)	TranscribersOfReddit (0.02%) redsox (0.04%)
Common misspellings	PewdiepieSubmissions (2.52%) AskOuija (1.79%)	MemeEconomy (0.06%) TranscribersOfReddit (0.15%)
G-dropping	teenagersnew (3.82%) Kanye (3.46%)	legaladvice (0.05%) changemyview (0.05%)
Vowel omission	magicTCG (3.50%) ClashRoyale (1.73%)	Christianity (0.01%) politics (0.01%)
Nearby character substitution	teenagersnew (0.06%) Drugs (0.03%)	TranscribersOfReddit (0.00%) teslamotors (0.00%)
British vs. American spelling	ukpolitics (94.88%) unitedkingdom (93.66%)	Dodgers (2.09%) NYYankees (2.37%)

Table 3: Spelling variation in Reddit, for each subreddit the fraction of non-conventional forms is shown.

4 Experiments

We now examine whether spelling variation is represented in a structured way in the embedding space using the selected pairs (Section 3.2). No intrinsic evaluation method is perfect—we therefore present three different types of analysis that complement each other: analogy tests (Section 4.2), a similarity analysis (Section 4.3), and diagnostic classifier tests (Section 4.4). Our goal is not to achieve the highest performance. Instead, by analyzing the performance—and how this differs between spelling variation types—we obtain new insights on what embeddings capture about spelling variation.

4.1 Word Embeddings

We experiment with two word embedding models, which map words to a d-dimensional space. The first is the continuous **skipgram** model with negative sampling (Mikolov et al., 2013a), using the gensim implementation (Řehůřek and Sojka, 2010). The skipgram model *does not* take the spelling of a word into account. As a natural comparison, we therefore also experiment with **fastText** (Bojanowski et al., 2017),[8] an extension of the continuous skipgram model, which *does* take spelling into account: each n-gram is represented by a vector representation and a word is represented by the sum of the vector representations of its n-grams. Both skipgram and fastText use the contexts of words to learn the embeddings. More specifically, for a given word w_i the surrounding words $\{w_{i-k}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+k}\}$ are used as context. *Although contextual factors associated with spelling variation are not taken into account explicitly, using the surrounding words as context could already implicitly capture such factors to some extent, causing spelling variation patterns to be encoded in the embeddings.* The embeddings were trained with a min. freq. count of 20, a window size of 5, and dimensions from 50–300, with a step size of 50. We used the default settings for the remaining parameters.

[8] https://fasttext.cc

4.2 Analogies

4.2.1 Experimental Setup

Analogies are frequently used to analyze whether word embeddings encode certain relations, by testing whether the right answers can be recovered using vector operations (Mikolov et al., 2013b; Levy and Goldberg, 2014). Here, the analogies are based on pairs of spelling variants. Take two pairs, e.g., (**a**: *cookin*, **a***: *cooking*) and (**b**: *movin*, **b***: *moving*). Then, the assumption is that the following holds: *cookin* − *cooking* ≈ *movin* − *moving*, or **a** − **a*** ≈ **b** − **b***. If **b*** is unknown, it can be found with:

$$\underset{\mathbf{b}^* \in V}{\operatorname{argmax}} \, cos(\mathbf{b}^*, \mathbf{b} - \mathbf{a} + \mathbf{a}^*)$$

The inputs **a**, **b** and **a*** are excluded as answers. This method is referred to by Levy and Goldberg (2014) as **3COSADD**. For example, given *cooking* - *cookin* + *movin*, words are ranked according to their cosine similarity with the resulting vector. The goal is to rank the correct answer (*moving*) at the top.

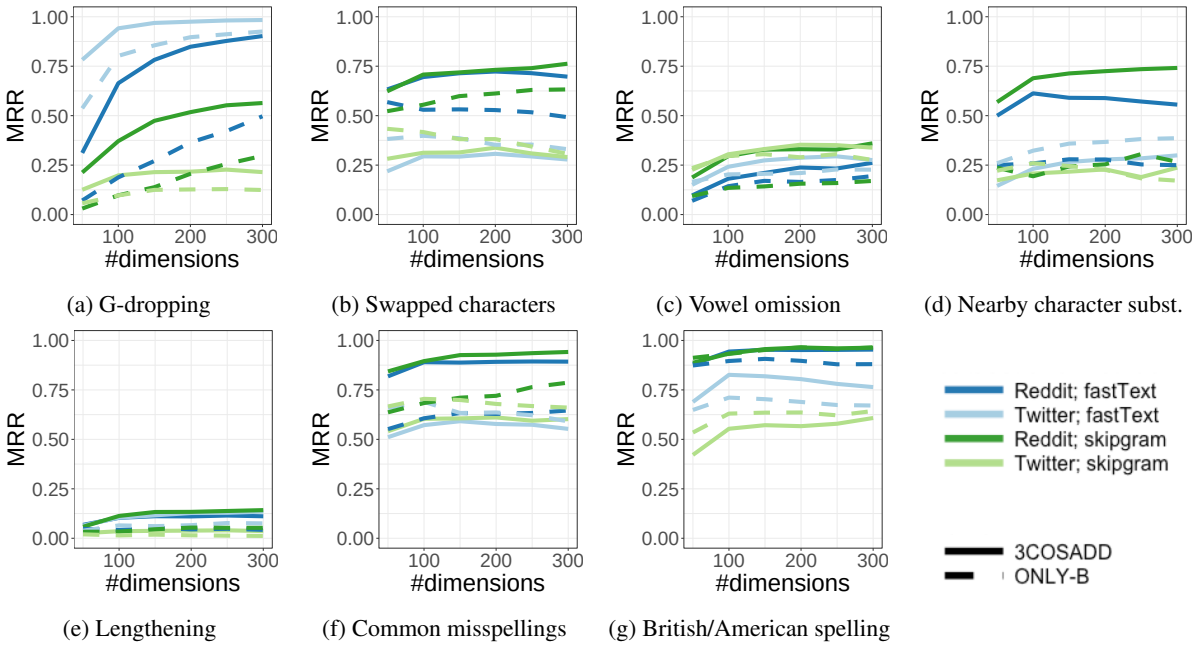
However, Linzen (2016) notes that results obtained this way can be misleading: the offsets are often very small, so that in practice often just the nearest neighbor to **b** is returned. We thus follow Linzen (2016) by reporting a control setting that just returns the nearest neighbor of **b** (**ONLY-B**), e.g., in our example returning the nearest neighbor of *movin*. A higher performance using the analogy setting (**3COSADD**) over the control setting (**ONLY-B**) would indicate that some relevant information is encoded in the embeddings.

We generate analogy instances by combining each pair (consisting of a conventional and non-conventional form) with a random other pair with the same spelling variation type. We do this ten times per pair, e.g., resulting in 2,580 analogy instances for lengthening. We generate analogies that aim to recover the *conventional* form. Pairs with a lengthened form are matched to pairs with the same amount of lengthening. We report accuracy for the best match and the mean reciprocal rank.

4.2.2 Results

The results are presented in Figure 2 (Mean Reciprocal Rank (MRR)) and Table 4 (accuracy).

3COSADD versus the control setting (ONLY-B) In most settings, the analogy setting (**3COSADD**) attains a higher performance than the control setting (**ONLY-B**).



(a) G-dropping (b) Swapped characters (c) Vowel omission (d) Nearby character subst.

(e) Lengthening (f) Common misspellings (g) British/American spelling

Figure 2: Analogy-based evaluation, reported using Mean Reciprocal Rank (MRR)

Spelling variation types The performance on g-dropping pairs is generally high and a strong improvement over the control setting (**ONLY-B**) is observed when using the analogy setup (**3COSADD**). Here, we also see a clear improvement in performance as the number of dimensions increases. FastText does particularly well, probably because the conventional forms can be recovered using fixed operations (e.g., adding a 'g').

In contrast, performance on the lengthening pairs is very low, with near zero accuracies. The control setting (**ONLY-B**) performs badly, and through manual inspection we found that both embedding models return many lengthened forms. This might be exacerbated with fastText, which seems to be sensitive to the n-grams in lengthened forms. However, even for lengthening we do observe a slight improvement from **3COSADD** over the control setting (**ONLY-B**), for both fastText and skipgram.

For the British/American spelling pairs we observe a clear performance difference between the two datasets. The performance on Twitter is much lower, but this is unsurprising as the Twitter dataset is constrained to the London area. The performance for the control setting (**ONLY-B**) on the swapped characters, common misspellings and British/American spelling pairs is relatively high. Our second analysis (Section 4.3) helps us to understand this trend.

data	fastText ONLY-B	fastText 3COSADD	skipgram ONLY-B	skipgram 3COSADD
g-dropping				
Reddit	0.21	0.67	0.10	0.38
Twitter	0.77	0.92	0.08	0.15
lengthening				
Reddit	0.01	0.04	0.02	0.08
Twitter	0.03	0.05	0.01	0.02
vowel omission				
Reddit	0.08	0.15	0.08	0.24
Twitter	0.11	0.18	0.20	0.25
British/American spelling				
Reddit	0.85	0.93	0.94	0.93
Twitter	0.48	0.66	0.53	0.47

data	fastText ONLY-B	fastText 3COSADD	skipgram ONLY-B	skipgram 3COSADD
swapped characters				
Reddit	0.41	0.60	0.52	0.66
Twitter	0.20	0.17	0.30	0.23
nearby character substitution				
Reddit	0.10	0.46	0.13	0.63
Twitter	0.23	0.17	0.14	0.14
common misspellings				
Reddit	0.48	0.83	0.57	0.88
Twitter	0.50	0.43	0.63	0.54

Table 4: Mean accuracies of analogy experiments (300 dim.)

FastText vs. skipgram The skipgram model does not take a word's spelling into account, while the fastText model does. FastText does particularly well on g-dropping (Figure 2a). This is expected, as the right answer for these cases can be recovered exactly using orthographic operations. However, the skipgram model's performance is notable as the analogy setting (**3COSADD**) improves over the control setting (**ONLY-B**) for most cases. Our results therefore indicate that skipgram, a spelling-agnostic embedding method, captures contextual factors that correlate with the use of different types of spelling variation. Moreover, in many settings it attains a higher performance than fastText. We note that we have not tuned the models' hyperparameters, and more in-depth comparisons are thus left for future work.

4.3 Word Similarity

To shed more light on the analogy results, we also compute the average cosine similarity between the embeddings of the non-conventional and conventional forms of each pair (Table 5).

With skipgram, we can broadly distinguish between two cases. We find that the types with the lowest average cosine similarities are the ones where the spelling variation is very likely *intentional*: g-dropping, lengthening and the omission of vowels. In contrast, the cosine similarity is higher for the types where the variation is probably often *unintentional*: swapping of characters and common misspellings. When spelling variation is intentional, we expect their usage to be highly context dependent (e.g., the *same* author might or might not use lengthening depending on the situation).

type	fastText	skipgram
g-dropping	0.72	0.48
lengthening	0.52	0.46
swapped characters	0.64	0.66
common misspellings	0.72	0.71
keyboard substitution	0.56	0.56
vowel omission	0.46	0.48
British vs. American	0.81	0.72

type	fastText	skipgram
random	0.14	0.12
BATS (verb inf-ing)	0.69	0.61
BATS (noun plural)	0.72	0.64

Table 5: Mean cosine similarities (300 dim.), averaged over both Twitter and Reddit embeddings

Different from these types of spelling variation are the British versus American spellings. Both forms are highly conventional and exhibit strong regional variation. Generally, we would expect less variation across social contexts and within individuals. For these pairs we also find high cosine similarities between the British and American spellings, for both fastText and skipgram.

Overall, fastText and skipgram show similar trends, except for g-dropping. These pairs tend to have a high cosine similarity with fastText embeddings. This is not surprising: the forms only differ in one character and fastText takes the n-grams into account. However, g-dropping tends to be used in specific social contexts, and fastText embeddings might represent the conventional and non-conventional forms artificially close to each other. Generally, the spelling variation types with high cosine similarity between the pairs are also the ones where just returning the nearest neighbor (**`ONLY-B`**) performs relatively well (Figure 2), for example, for common misspellings.

To test whether the type of spelling variation indeed has a significant effect on the cosine similarity between the conventional and non-conventional forms, we fit a logistic regression model with the cosine similarity as the dependent variable. As the independent variables we include the spelling variation type, dataset, embedding model, and corpus frequencies (for both the conventional and non-conventional form). The spelling variation types remain highly significant. The model achieves a fit of R^2=0.37.

As reference points (Table 5), we report cosine similarities for random pairs and pairs from two linguistic relations from BATS (Gladkova et al., 2016), verb infinitive:participle (*ask*:*asking*) and noun:plural (*car*:*cars*). The cosine similarities are low for the random pairs. For the BATS pairs, the similarities are higher, but still lower or comparable to common misspellings and British/American spelling pairs.

4.4 Diagnostic Classifiers

As highlighted by Linzen (2016) and Rogers et al. (2017), analogy experiments are limited as linguistic properties may be encoded in embeddings in ways not visible through a linear offset. Following Hupkes et al. (2018), Adi et al. (2017), Zhu et al. (2018), and others, we perform classification experiments using so-called diagnostic or probing classifiers. We build classifiers to predict the type of spelling variation based on the word embeddings alone, building on the core assumption that their performance reflects the extent to which information about spelling variation is encoded in the embeddings.

4.4.1 Experimental Setup

We frame the task as a multi-class classification problem: for a given pair (consisting of a non-conventional and conventional form), predict the correct type of spelling variation of the non-conventional form. For example, for the pair *thaaaaanks* and *thanks*, the correct output is the lengthening class. In total, we have 794 pairs and seven classes. We use logistic regression with L2 regularization implemented using scikit-learn (Pedregosa et al., 2011), opting deliberately for a linear classifier. No parameter tuning was performed and results are reported using ten-fold cross validation. With the analogy experiments, a higher number of dimensions generally led to better performance. We therefore only experiment with 300 dimensions. We experiment with both normalized and unnormalized embeddings.

Control settings A challenge is that certain types of spelling variation are associated with certain types of words (similar to problems highlighted by Levy et al. (2015)). For example, in our data *g*-dropping occurs mostly with verbs. Thus, a classifier might just learn to recognize, say, verbs to identify *g*-dropping

setting	dataset	embedding	norm.	CONTROL1 (F1) micro	CONTROL1 (F1) macro	CONTROL2 (F1) micro	CONTROL2 (F1) macro	DIFF-EMB (F1) micro	DIFF-EMB (F1) macro
1	Reddit	fastText	yes	0.67	0.41	0.72	0.48	0.89	0.74
2	Reddit	fastText	no	0.67	0.47	0.73	0.54	0.90	0.79
3	Reddit	skipgram	yes	0.66	0.41	0.72	0.48	0.87	0.74
4	Reddit	skipgram	no	0.64	0.45	0.71	0.53	0.88	0.78
5	Twitter	fastText	yes	0.69	0.46	0.73	0.54	0.79	0.61
6	Twitter	fastText	no	0.68	0.50	0.71	0.55	0.80	0.67
7	Twitter	skipgram	yes	0.70	0.47	0.72	0.53	0.73	0.57
8	Twitter	skipgram	no	0.66	0.49	0.70	0.56	0.70	0.58

Table 6: Diagnostic classifier results, with norm. indicating whether embeddings are normalized.

instead of the real phenomenon. We therefore include a control setting (**CONTROL1**), in which a classifier predicts the spelling variation type based on the embeddings of the conventional form alone.

Moreover, word embeddings encode information about word frequency (Schnabel et al., 2015), which can correlate with the type of spelling variation. For example, in our data, the g-dropping forms have higher corpus frequencies than the character substitution forms. We therefore have another control setting that extends the previous one with a feature indicating the frequency of the non-conventional form (on a log scale), **CONTROL2**. A classifier being able to attain better performance than these control settings is evidence that embeddings encode information associated with the type of spelling variation.

Main setting Our main classifier is **DIFF-EMB**, which predicts the type of spelling variation based on the conventional form's embedding subtracted from the non-conventional form's embedding.

4.4.2 Results

Table 6 shows both macro and micro F1 scores. We see a clear performance improvement of **DIFF-EMB** over the control settings.

Control settings **CONTROL1** reaches relatively high F1 scores, even though its predictions *are not based* on the embeddings of the non-conventional forms, because many types of spelling variation tend to occur with certain word categories. For example, the classifier using normalized fastText embeddings trained with Twitter data (setting 5) obtains a high F-score (0.91) for g-dropping. It also reaches good performance for British/American spelling pairs (F1 score: 0.72) and for lengthening (F1 score: 0.75). In our data, for example, lengthening occurs mostly on interjections and on highly conversational word forms, as well as words associated with expressing information about quality and quantity. Low performance was obtained for the vowel omission (F1 score: 0.04) and keyboard substitution (F1 score: 0.00) pairs. These two types also have fewer instances in our data. We also see a consistent but small improvement with **CONTROL2**, which adds frequency information. For both **CONTROL1** and **CONTROL2**, unnormalized embeddings perform slightly better than normalized embeddings on macro F1 scores. Finally, we note that the results are considerably higher than what a majority-class classifier, which would always predict lengthening, achieves (F1 0.32 micro, 0.07 macro).

Main setting **DIFF-EMB** achieves a clear performance improvement over the control settings, which is evidence that the embeddings capture useful information that go beyond information associated with the conventional form. Performance using unnormalized embeddings is consistently better, one possible reason is that the norm of an embedding is related to the corpus frequency of the form. Better performance is also achieved with Reddit embeddings. Furthermore, fastText embeddings perform slightly better than skipgram embeddings, which is expected given that fastText takes the spelling of words into account.

On the Reddit data, both (normalized) fastText and skipgram embeddings (settings 1 and 3) lead to high performance (F1-score$\geq$0.90) for British/American pairs, g-dropping, lengthening, and common misspellings. Performance was low for keyboard substitution pairs, which were often misclassified as having swapped characters.

5 Conclusion

Deviations from conventional spelling norms often carry social meaning (Sebba, 2007). Depending on the application, it may thus not be desirable to map all spelling variants to the same point in the embedding space, especially in the areas of computational social science (Lazer et al., 2009) and computational sociolinguistics (Nguyen et al., 2016). This paper is a first step towards analyzing how spelling variation is captured in word embeddings. We presented three analyses in which we looked at seven types of spelling variation. Taken together, we found that even embeddings trained using the skipgram model—which does not take spelling into account—encode spelling variation patterns to some extent. Moreover, our analysis of the cosine similarities between non-conventional and conventional forms for different types of spelling variation suggest a link between the intentionality of the variation and the distance to the conventional forms. A possible explanation is that intentional spelling variation might occur in more specific contexts; future work should investigate this more deeply. We also note that although non-conventional spellings generally have the same referential meaning as their conventional form, they can sometimes also mark specific new meanings of words (Grieve et al., 2017). As future work, contextual embeddings (Devlin et al., 2019; Peters et al., 2018) and other spelling variation types can be explored. We also plan to investigate other classifiers and feature representations to capture the interaction between two embeddings for our diagnostic classifiers (Levy et al., 2015).

Acknowledgements

DN is supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the Alan Turing Institute Fellowship (TU/A/000006). This work is part of the research programme Veni with project number VI.Veni.192.130, which is (partly) financed by the Dutch Research Council (NWO).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR 2017*.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570.

Kathryn Campbell-Kibler. 2006. *Listener perceptions of sociolinguistic variables: The case of (ING)*. Ph.D. thesis, Stanford University, 07.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $\$\&!\#*$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*, 21(1):99–127.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61(1):907–926.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723.

Erez Levon and Sue Fox. 2014. Social salience and the sociolinguistic monitor: A case study of ing and th-fronting in Britain. *Journal of English Linguistics*, 42(3):185–217.

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.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado, May–June.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 439–448.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: exploratory search and topic summarization for Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 384–385.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Deana Pennell and Yang Liu. 2011. A character-level machine translation approach for normalization of SMS abbreviations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 974–982, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Aleksandra Piktus, Necati Bora Edizel, Piotr Bojanowski, Edouard Grave, Rui Ferreira, and Fabrizio Silvestri. 2019. Misspelling oblivious word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3226–3234.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Mark Sebba. 2007. *Spelling and society: The culture and politics of orthography around the world*. Cambridge University Press.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2018. Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 1–6.

Timothy Francis John Shortis. 2016. *Orthographic practices in SMS text messaging as a case signifying diachronic change in linguistic and semiotic resources*. Ph.D. thesis, UCL (University College London).

Caroline Tagg. 2009. *A corpus linguistics study of SMS text messaging*. Ph.D. thesis, University of Birmingham.

Meredith Tamminga. 2017. Matched guise effects can be robust to speech style. *The Journal of the Acoustical Society of America*, 142(1):EL18–EL23.

Rachael Tatman. 2015. #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle*, 25(2):97–108.

Crispin Thurlow and Alex Brown. 2003. Generation txt? The sociolinguistics of young people's text-messaging. *Discourse analysis online*.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.

Laurens J.P. van der Maaten and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9(nov):2579–2605.

Zijian Wang, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *Proceedings of the 2019 World Wide Web Conference*, pages 2056–2067.

Xunjie Zhu, Tingfeng Li, and Gerard Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637.