An Unsupervised Model for Text Message Normalization

Paul Cook
Department of Computer Science
University of Toronto
Toronto, Canada
pcook@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

Cell phone text messaging users express themselves briefly and colloquially using a variety of creative forms. We analyze a sample of creative, non-standard text message word forms to determine frequent word formation processes in texting language. Drawing on these observations, we construct an unsupervised noisy-channel model for text message normalization. On a test set of 303 text message forms that differ from their standard form, our model achieves 59% accuracy, which is on par with the best supervised results reported on this dataset.

1 Text Messaging

Cell phone text messages—or SMS—contain many shortened and non-standard forms due to a variety of factors, particularly the desire for rapid text entry (Grinter and Eldridge, 2001; Thurlow, 2003).[1] Furthermore, text messages are written in an informal register; non-standard forms are used to reflect this, and even for personal style (Thurlow, 2003). These factors result in tremendous linguistic creativity, and hence many novel lexical items, in the language of text messaging, or *texting language*.

Normalization of non-standard forms—converting non-standard forms to their standard forms—is a challenge that must be tackled before other types of natural language processing can take place (Sproat et al., 2001). In the case of text messages, text-to-speech synthesis may be particularly useful for the visually impaired; automatic translation has also been considered (e.g., Aw et al., 2006). For texting language, given the abundance of creative forms, and the wide-ranging possibilities for creating new forms, normalization is a particularly important problem, and has indeed received some attention in computational linguistics (e.g., Aw et al., 2006; Choudhury et al., 2007; Kobus et al., 2008).

In this paper we propose an unsupervised noisy channel method for texting language normalization, that gives performance on par with that of a supervised system. We pursue unsupervised approaches to this problem, as large collections of text messages, and their corresponding standard forms, are not readily available.[2] Furthermore, other forms of computer-mediated communication, such as Internet messaging, exhibit creative phenomena similar to text messaging, although at a lower frequency (Ling and Baron, 2007). Moreover, technological changes, such as new input devices, are likely to have an impact on the language of such media (Thurlow, 2003).[3] An unsupervised approach, drawing on linguistic properties of creative word formations, has the potential to be adapted for normalization of text in other similar genres—such as Internet discussion forums—without the cost of developing a large training corpus. Moreover, normalization may be particularly important for such genres, given the

[1] The number of characters in a text message may also be limited to 160 characters, although this is not always the case.

[2] One notable exception is Fairon and Paumier (2006), although this resource is in French. The resource used in our study, Choudhury et al. (2007), is quite small in comparison.

[3] The rise of other technology, such as word prediction, could reduce the use of abbreviations, although it's not clear such technology is widely used (Grinter and Eldridge, 2001).

Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, pages 71–78,
Boulder, Colorado, June 2009.

Formation type	Freq.	Example
Stylistic variation	152	*betta* (*better*)
Subseq. abbrev.	111	*dng* (*doing*)
Prefix clipping	24	*hol* (*holiday*)
Syll. letter/digit	19	*neway* (*anyway*)
G-clipping	14	*talkin* (*talking*)
Phonetic abbrev.	12	*cuz* (*because*)
H-clipping	10	*ello* (*hello*)
Spelling error	5	*darliog* (*darling*)
Suffix clipping	4	*morrow* (*tomorrow*)
Punctuation	3	*b/day* (*birthday*)
Unclear	34	*mobs* (*mobile*)
Error	12	*gal* (**girl*)
Total	400	

Table 1: Frequency of texting forms in the development set by formation type.

need for applications such as translation and question answering.

We observe that many creative texting forms are the result of a small number of specific word formation processes. Rather than using a generic error model to capture all of them, we propose a mixture model in which each word formation process is modeled explicitly according to linguistic observations specific to that formation.

2 Analysis of Texting Forms

To better understand the creative processes present in texting language, we categorize the word formation process of each texting form in our development data, which consists of 400 texting forms paired with their standard forms.[4] Several iterations of categorization were done in order to determine sensible categories, and ensure categories were used consistently. Since this data is only to be used to guide the construction of our system, and not for formal evaluation, only one judge (a native English speaking author of this paper) categorized the expressions. The findings are presented in Table 1.

Stylistic variations, by far the most frequent category, exhibit non-standard spelling, such as representing sounds phonetically. Subsequence abbreviations, also very frequent, are composed of a subsequence of the graphemes in a standard form, often omitting vowels. These two formation types account for approximately 66% of our development data; the remaining formation types are much less frequent. Prefix clippings and suffix clippings consist of a prefix or suffix, respectively, of a standard form, and in some cases a diminutive ending; we also consider clippings which omit just a *g* or *h* from a standard form as they are rather frequent.[5] A single letter or digit can be used to represent a syllable; we refer to these as syllabic (syll.) letter/digit. Phonetic abbreviations are variants of clippings and subsequence abbreviations where some sounds in the standard form are represented phonetically. Several texting forms appear to be spelling errors; we took the layout of letters on cell phone keypads into account when making this judgement. The items that did not fit within the above texting form categories were marked as unclear. Finally, for some expressions the given standard form did not appear to be appropriate. For example, *girl* is not the standard form for the texting form *gal*; rather, *gal* is an English word that is a colloquial form of *girl*. Such cases were marked as errors.

No texting forms in our development data correspond to multiple standard form words, e.g., *wanna* for *want to*.[6] Since such forms are not present in our development data, we assume that a texting form always corresponds to a single standard form word.

It is important to note that some text forms have properties of multiple categories, e.g., *bak* (*back*) could be considered a stylistic variation or a subsequence abbreviation. In such cases, we simply attempt to assign the most appropriate category.

The design of our model for text message normalization, presented below, uses properties of the observed formation processes.

3 An Unsupervised Noisy Channel Model for Text Message Normalization

Let S be a sentence consisting of *standard forms* $s_1 s_2 ... s_n$; in this study the standard forms are reg-

[4] Most texting forms have a unique standard form; however, some have multiple standard forms, e.g., *will* and *well* can both be shortened to *wl*. In such cases we choose the category of the most frequent standard form; in the case of frequency ties we choose arbitrarily among the categories of the standard forms.

[5] Thurlow (2003) also observes an abundance of g-clippings.

[6] A small number of similar forms, however, appear with a single standard form word, and are therefore marked as errors.

ular English words. Let T be a sequence of *texting forms* $t_1 t_2 ... t_n$, which are the texting language realization of the standard forms, and may differ from the standard forms. Given a sequence of texting forms T, the challenge is then to determine the corresponding standard forms S.

Following Choudhury et al. (2007)—and various approaches to spelling error correction, such as, e.g., Mays et al. (1991)—we model text message normalization using a noisy channel. We want to find $\text{argmax}_S P(S|T)$. We apply Bayes rule and ignore the constant term $P(T)$, giving $\text{argmax}_S P(T|S)P(S)$. Making the independence assumption that each t_i depends only on s_i, and not on the context in which it occurs, as in Choudhury et al., we express $P(T|S)$ as a product of probabilities: $\text{argmax}_S \left(\prod_i P(t_i|s_i)\right) P(S)$.

We note in Section 2 that many texting forms are created through a small number of specific word formation processes. Rather than model each of these processes at once using a generic model for $P(t_i|s_i)$, as in Choudhury et al., we instead create several such models, each corresponding to one of the observed common word formation processes. We therefore rewrite $P(t_i|s_i)$ as $\sum_{wf} P(t_i|s_i, wf)P(wf)$ where wf is a word formation process, e.g., subsequence abbreviation. Since, like Choudhury et al., we focus on the word model, we simplify our model as below.

$$\text{argmax}_{s_i} \sum_{wf} P(t_i|s_i, wf)P(wf)P(s_i)$$

We next explain the components of the model, $P(t_i|s_i, wf)$, $P(wf)$, and $P(s_i)$, referred to as the word model, word formation prior, and language model, respectively.

3.1 Word Models

We now consider which of the word formation processes discussed in Section 2 should be captured with a word model $P(t_i|s_i, wf)$. We model stylistic variations and subsequence abbreviations simply due to their frequency. We also choose to model prefix clippings since this word formation process is common outside of text messaging (Kreidler, 1979; Algeo, 1991) and fairly frequent in our data. Although g-clippings and h-clippings are moderately frequent, we do not model them, as these very specific word formations are also (non-prototypical) subsequence abbreviations. We do not model syllabic letters and digits, or punctuation, explicitly; instead, we simply substitute digits with a graphemic representation (e.g., *4* is replaced by *for*), and remove punctuation, before applying the model. The other less frequent formations—phonetic abbreviations, spelling errors, and suffix clippings—are not modeled; we hypothesize that the similarity of these formation processes to those we do model will allow the system to perform reasonably well on them.

graphemes	w	i	th	ou	t
phonemes	w	ɪ	θ	au	t

Table 2: Grapheme–phoneme alignment for *without*.

3.1.1 Stylistic Variations

We propose a probabilistic version of edit-distance—referred to here as edit-probability—inspired by Brill and Moore (2000) to model $P(t_i|s_i, \text{stylistic variation})$. To compute edit-probability, we consider the probability of each edit operation—substitution, insertion, and deletion—instead of its cost, as in edit-distance. We then simply multiply the probabilities of edits as opposed to summing their costs.

In this version of edit-probability, we allow two-character edits. Ideally, we would compute the edit-probability of two strings as the sum of the edit-probability of each partitioning of those strings into one or two character segments. However, following Brill and Moore, we approximate this by the probability of the partition with maximum probability. This allows us to compute edit-probability using a simple adaptation of edit-distance, in which we consider edit operations spanning two characters at each cell in the chart maintained by the algorithm.

We then estimate two probabilities: $P(g_t|g_s, pos)$ is the probability of texting form grapheme g_t given standard form grapheme g_s at position pos, where pos is the beginning, middle, or end of the word; $P(h_t|p_s, h_s, pos)$ is the probability of texting form graphemes h_t given the standard form phonemes p_s and graphemes h_s at position pos. h_t, p_s, and h_s can be a single grapheme or phoneme, or a bigram.

We compute edit-probability between the graphemes of s_i and t_i. When filling each cell in the chart, we consider edit operations between

segments of s_i and t_i of length 0–2, referred to as a and b, respectively. If a aligns with phonemes in s_i, we also consider those phonemes, p. In our lexicon, the graphemes and phonemes of each word are aligned according to the method of Jiampojamarn et al. (2007). For example, the alignment for *without* is given in Table 2. The probability of each edit operation is then determined by three properties—the length of a, whether a aligns with any phonemes in s_i, and if so, p—as shown below:

$|a|$= 0 or 1, not aligned w/ s_i phonemes: $P(b|a, pos)$

$|a|$= 2, not aligned w/ s_i phonemes: 0

$|a|$= 1 or 2, aligned w/ s_i phonemes: $P(b|p, a, pos)$

3.1.2 Subsequence Abbreviations

We model subsequence abbreviations according to the equation below:

$$P(t_i|s_i, \text{subseq abrv}) = \begin{cases} c & \text{if } t_i \text{ is a subseq of } s_i \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant.

Note that this is similar to the error model for spelling correction presented by Mays et al. (1991), in which all words (in our terms, all s_i) within a specified edit-distance of the out-of-vocabulary word (t_i in our model) are given equal probability. The key difference is that in our formulation, we only consider standard forms for which the texting form is potentially a subsequence abbreviation.

In combination with the language model, $P(t_i|s_i, \text{subseq abbrev})$ assigns a non-zero probability to each standard form s_i for which t_i is a subsequence, according to the likelihood of s_i (under the language model). The models interact in this way since we expect a standard form to be recognizable relative to the other words for which t_i could be a subsequence abbreviation

3.1.3 Prefix Clippings

We model prefix clippings similarly to subsequence abbreviations.

$$P(t_i|s_i, \text{prefix clipping}) = \begin{cases} c & \text{if } t_i \text{ is possible} \\ & \text{pre. clip. of } s_i \\ 0 & \text{otherwise} \end{cases}$$

Kreidler (1979) observes that clippings tend to be mono-syllabic and end in a consonant. Furthermore, when they do end in a vowel, it is often of a regular form, such as *telly* for *television* and *breaky* for *breakfast*. We therefore only consider $P(t_i|s_i, \text{prefix clipping})$ if t_i is a prefix clipping according to the following heuristics: t_i is mono-syllabic after stripping any word-final vowels, and subsequently removing duplicated word-final consonants (e.g, *telly* becomes *tel*, which is a candidate prefix clipping). If t_i is not a prefix clipping according to these criteria, $P(t_i|s_i)$ simply sums over all models except prefix clipping.

3.2 Word Formation Prior

Keeping with our goal of an unsupervised method, we estimate $P(wf)$ with a uniform distribution. We also consider estimating $P(wf)$ using maximum likelihood estimates (MLEs) from our observations in Section 2. This gives a model that is not fully unsupervised, since it relies on labelled training data. However, we consider this a lightly-supervised method, since it only requires an estimate of the frequency of the relevant word formation types, and not labelled texting form–standard form pairs.

3.3 Language Model

Choudhury et al. (2007) find that using a bigram language model estimated over a balanced corpus of English had a negative effect on their results compared with a unigram language model, which they attribute to the unique characteristics of text messaging that were not reflected in the corpus. We therefore use a unigram language model for $P(s_i)$, which also enables comparison with their results. Nevertheless, alternative language models, such as higher order ngram models, could easily be used in place of our unigram language model.

4 Materials and Methods

4.1 Datasets

We use the data provided by Choudhury et al. (2007) which consists of texting forms—extracted from a collection of 900 text messages—and their manually determined standard forms. Our development data—used for model development and discussed in Section 2—consists of the 400 texting form types that are not in Choudhury et al.'s held-out test set, and that are not the same as one of their standard

forms. The test data consists of 1213 texting forms and their corresponding standard forms. A subset of 303 of these texting forms differ from their standard form.[7] This subset is the focus of this study, but we also report results on the full dataset.

4.2 Lexicon

We construct a lexicon of potential standard forms such that it contains most words that we expect to encounter in text messages, yet is not so large as to make it difficult to identify the correct standard form. Our subjective analysis of the standard forms in the development data is that they are frequent, non-specialized, words. To reflect this observation, we create a lexicon consisting of all single-word entries containing only alphabetic characters found in both the CELEX Lexical Database (Baayen et al., 1995) and the CMU Pronouncing Dictionary.[8] We remove all words of length one (except *a* and *I*) to avoid choosing, e.g., the letter *r* as the standard form for the texting form *r*. We further limit the lexicon to words in the 20K most frequent alphabetic unigrams, ignoring case, in the Web 1T 5-gram Corpus (Brants and Franz, 2006). The resulting lexicon contains approximately 14K words, and excludes only three of the standard forms—*cannot*, *email*, and *online*—for the 400 development texting forms.

4.3 Model Parameter Estimation

MLEs for $P(g_t|g_s, pos)$—needed to estimate $P(t_i|s_i, \text{stylistic variation})$—could be estimated from texting form–standard form pairs. However, since our system is unsupervised, no such data is available. We therefore assume that many texting forms, and other similar creative shortenings, occur on the web. We develop a number of character substitution rules, e.g., $s \Rightarrow z$, and use them to create hypothetical texting forms from standard words. We then compute MLEs for $P(g_t|g_s, pos)$ using the frequencies of these derived forms on the web.

We create the substitution rules by examining examples in the development data, considering fast speech variants and dialectal differences (e.g., voicing), and drawing on our intuition. The derived forms are produced by applying the substitution rules to the words in our lexicon. To avoid considering forms that are themselves words, we eliminate any form found in a list of approximately 480K words taken from SOWPODS[9] and the Moby Word Lists.[10] Finally, we obtain the frequency of the derived forms from the Web 1T 5-gram Corpus.

To estimate $P(h_t|p_s, h_s, pos)$, we first estimate two simpler distributions: $P(h_t|h_s, pos)$ and $P(h_t|p_s, pos)$. $P(h_t|h_s, pos)$ is estimated in the same manner as $P(g_t|g_s, pos)$, except that two character substitutions are allowed. $P(h_t|p_s, pos)$ is estimated from the frequency of p_s, and its alignment with h_t, in a version of CELEX in which the graphemic and phonemic representation of each word is many–many aligned using the method of Jiampojamarn et al. (2007).[11] $P(h_t|p_s, h_s, pos)$ is then an evenly-weighted linear combination of $P(h_t|h_s, pos)$ and $P(h_t|p_s, pos)$. Finally, we smooth each of $P(g_t|g_s, pos)$ and $P(h_t|p_s, h_s, pos)$ using add-alpha smoothing.

We set the constant c in our word models for subsequence abbreviations and prefix clippings such that $\sum_{s_i} P(t_i|s_i, wf)P(s_i) = 1$. We similarly normalize $P(t_i|s_i, \text{stylistic variation})P(s_i)$.

We use the frequency of unigrams (ignoring case) in the Web 1T 5-gram Corpus to estimate our language model. We expect the language of text messaging to be more similar to that found on the web than that in a balanced corpus of English.

4.4 Evaluation Metrics

To evaluate our system, we consider three accuracy metrics: in-top-1, in-top-10, and in-top-20.[12] In-top-n considers the system correct if a correct standard form is in the n most probable standard forms. The in-top-1 accuracy shows how well the system determines the correct standard form; the in-top-10

[7] Choudhury et al. report that this dataset contains 1228 texting forms. We found it to contain 1213 texting forms corresponding to 1228 standard forms (recall that a texting form may have multiple standard forms). There were similar inconsistencies with the subset of texting forms that differ from their standard forms. Nevertheless, we do not expect these small differences to have an appreciable effect on the results.

[8] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[9] `http://en.wikipedia.org/wiki/SOWPODS`

[10] `http://icon.shef.ac.uk/Moby/`

[11] We are very grateful to Sittichai Jiampojamarn for providing this alignment.

[12] These are the same metrics used by Choudhury et al. (2007), although we refer to them by different names.

Model	% accuracy		
	Top-1	Top-10	Top-20
Uniform	59.4	83.8	87.8
MLE	55.4	84.2	86.5
Choudhury et al.	59.9	84.3	88.7

Table 3: % in-top-1, in-top-10, and in-top-20 accuracy on test data using both estimates for $P(wf)$. The results reported by Choudhury et al. (2007) are also shown.

and in-top-20 accuracies may be indicative of the usefulness of the output of our system in other tasks which could exploit a ranked list of standard forms, such as machine translation.

5 Results and Discussion

In Table 3 we report the results of our system using both the uniform estimate and the MLE of $P(wf)$. Note that there is no meaningful random baseline to compare against here; randomly ordering the 14K words in our lexicon gives very low accuracy. The results using the uniform estimate of $P(wf)$—a fully unsupervised system—are very similar to the supervised results of Choudhury et al. (2007). Surprisingly, when we estimate $P(wf)$ using MLEs from the development data—resulting in a lightly-supervised system—the results are slightly worse than when using the uniform estimate of this probability. Moreover, we observe the same trend on development data where we expect to have an accurate estimate of $P(wf)$ (results not shown). We hypothesize that the ambiguity of the categories of text forms (see Section 2) results in poor MLEs for $P(wf)$, thus making a uniform distribution, and hence fully-unsupervised approach, more appropriate.

Results by Formation Type We now consider in-top-1 accuracy for each word formation type, in Table 4. We show results for the same word formation processes as in Table 1, except for h-clippings and punctuation, as no words of these categories are present in the test data. We present results using the same experimental setup as before with a uniform estimate of $P(wf)$ (All), and using just the model corresponding to the word formation process (Specific), where applicable.[13]

[13]In this case our model then becomes, for each word formation process wf, $\text{argmax}_{s_i} P(t_i|s_i, wf)P(s_i)$.

Formation type	Freq.	% in-top-1 acc.	
	$n = 303$	Specific	All
Stylistic variation	121	62.8	67.8
Subseq. abbrev.	65	56.9	46.2
Prefix clipping	25	44.0	20.0
G-clipping	56	-	91.1
Syll. letter/digit	16	-	50.0
Unclear	12	-	0.0
Spelling error	5	-	80.0
Suffix clipping	1	-	0.0
Phonetic abbrev.	1	-	0.0
Error	1	-	0.0

Table 4: Frequency (Freq.), and % in-top-1 accuracy using the formation-specific model where applicable (Specific) and all models (All) with a uniform estimate for $P(wf)$, presented by formation type.

We first examine the top panel of Table 3 where we compare the performance on each word formation type for both experimental conditions (Specific and All). We first note that the performance using the formation-specific model on subsequence abbreviations and prefix clippings is better than that of the overall model. This is unsurprising since we expect that when we know a texting form's formation process, and invoke a corresponding specific model, our system should outperform a model designed to handle a range of formation types. However, this is not the case for stylistic variations; here the overall model performs better than the specific model. We observed in Section 2 that some texting forms do not fit neatly into our categorization scheme; indeed, many stylistic variations are also analyzable as subsequence abbreviations. Therefore, the subsequence abbreviation model may benefit normalization of stylistic variations. This model, used in isolation on stylistic variations, gives an in-top-1 accuracy of 33.1%, indicating that this may be the case.

Comparing the performance of the individual word models on only word types that they were designed for (column Specific in Table 4), we see that the prefix clipping model is by far the lowest, indicating that in the future we should consider ways of improving this word model. One possibility is to incorporate phonemic knowledge. For example, both *friday* and *friend* have the same probability un-

der $P(t_i|s_i, \text{prefix clipping})$ for the texting form *fri*, which has the standard form *friday* in our data. (The language model, however, does distinguish between these forms.) However, if we consider the phonemic representations of these words, *friday* might emerge as more likely. Syllable structure information may also be useful, as we hypothesize that clippings will tend to be formed by truncating a word at a syllable boundary. We may similarly be able to improve our estimate of $P(t_i|s_i, \text{subseq. abrrev.})$. For example, both *text* and *taxation* have the same probability under this distribution, but intuitively *text*, the correct standard form in our data, seems more likely. We could incorporate knowledge about the likelihood of omitting specific characters, as in Choudhury et al. (2007), to improve this estimate.

We now examine the lower panel of Table 4, in which we consider the performance of the overall model on the word formation types that are not explicitly modeled. The very high accuracy on g-clippings indicates that since these forms are also a type of subsequence abbreviation, we do not need to construct a separate model for them. We in fact also conducted experiments in which g-clippings and h-clippings were modeled explicitly, but found these extra models to have little effect on the results.

Recall from Section 3.1 our hypothesis that suffix clippings, spelling errors, and phonetic abbreviations have common properties with formation types that we do model, and therefore the system will perform reasonably well on them. Here we find preliminary evidence to support this hypothesis as the accuracy on these three word formation types (combined) is 57.1%. However, we must interpret this result cautiously as it only considers seven expressions. On the syllabic letter and digit texting forms the accuracy is 50.0%, indicating that our heuristic to replace digits in texting forms with an orthographic representation is reasonable.

The performance on types of expressions that we did not consider when designing the system—unclear and error—is very poor. However, this has little impact on the overall performance as these expressions are rather infrequent.

Results by Model We now consider in-top-1 accuracy using each model on the 303 test expressions; results are shown in Table 5. No model on its own gives results comparable to those of the overall model (59.4%, see Table 3). This indicates that the overall model successfully combines information from the specific word formation models.

Model	% in-top-1 accuracy
Stylistic variation	51.8
Subseq. Abbrev.	44.2
Prefix clipping	10.6

Table 5: % in-top-1 accuracy on the 303 test expressions using each model individually.

Each model used on its own gives an accuracy greater than the proportion of expressions of the word formation type for which the model was designed (compare accuracies in Table 5 to the number of expressions of the corresponding word formation type in the test data in Table 4). As we note in Section 2, the distinctions between the word formation types are not sharp; these results show that the shared properties of word formation types enable a model for a specific formation type to infer the standard form of texting forms of other formation types.

All Unseen Data Until now we have discussed results on our test data of 303 texting forms which differ from their standard forms. We now consider the performance of our system on all 1213 unseen texting forms, 910 of which are identical to their standard form. Since our model was not designed with such expressions in mind, we slightly adapt it for this new task; if t_i is in our lexicon, we return that form as s_i, otherwise we apply our model as usual, using the uniform estimate of $P(wf)$. This gives an in-top-1 accuracy of 88.2%, which is very similar to the results of Choudhury et al. (2007) on this data of 89.1%. Note, however, that Choudhury et al. only report results on this dataset using a uniform language model;[14] since we use a unigram language model, it is difficult to draw firm conclusions about the performance of our system relative to theirs.

6 Related Work

Aw et al. (2006) model text message normalization as translation from the texting language into the

[14] Choudhury et al. do use a unigram language model for their experiments on the 303 texting forms which differ from their standard forms (see Section 3.3).

standard language. Kobus et al. (2008) incorporate ideas from both machine translation and automatic speech recognition for text message normalization. However, both of these approaches are supervised, and have only limited means for normalizing texting forms that do not occur in the training data.

Our work, like that of Choudhury et al. (2007), can be viewed as a noisy-channel model for spelling error correction (e.g., Mays et al., 1991; Brill and Moore, 2000), in which texting forms are seen as a kind of spelling error. Furthermore, like our approach to text message normalization, approaches to spelling correction have incorporated phonemic information (Toutanova and Moore, 2002).

The word model of the supervised approach of Choudhury et al. consists of hidden Markov models, which capture properties of texting language similar to those of our stylistic variation model. We propose multiple word models—corresponding to frequent texting language formation processes—and an unsupervised method for parameter estimation.

7 Conclusions

We analyze a sample of texting forms to determine frequent word formation processes in creative texting language. Drawing on these observations, we construct an unsupervised noisy-channel model for text message normalization. On an unseen test set of 303 texting forms that differ from their standard form, our model achieves 59% accuracy, which is on par with that obtained by the supervised approach of Choudhury et al. (2007) on the same data.

More research is required to determine the impact of our normalization method on the performance of a system that further processes the resulting text. In the future, we intend to improve our word models by incorporating additional linguistic knowledge, such as information about syllable structure. Since context likely plays a role in human interpretation of texting forms, we also intend to examine the performance of higher order ngram language models.

Acknowledgements

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the University of Toronto, and the Dictionary Society of North America.

References

John Algeo, editor. 1991. *Fifty Years Among the New Words*. Cambridge University Press, Cambridge.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40. Sydney.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX Lexical Database (release 2). Linguistic Data Consortium, University of Pennsylvania.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Corpus version 1.1.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of ACL 2000*, pages 286–293. Hong Kong.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition*, 10(3/4):157–174.

Cédrick Fairon and Sébastien Paumier. 2006. A translated corpus of 30,000 French SMS. In *Proceedings of LREC 2006*. Genoa, Italy.

Rebecca E. Grinter and Margery A. Eldridge. 2001. y do tngrs luv 2 txt msg. In *Proceedings of the 7th European Conf. on Computer-Supported Cooperative Work (ECSCW '01)*, pages 219–238. Bonn, Germany.

Sittichai Jiampojamarn, Gregorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proc. of NAACL-HLT 2007*, pages 372–379. Rochester, NY.

Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one? In *Proc. of the 22nd Int. Conf. on Computational Linguistics*, pp. 441–448. Manchester.

Charles W. Kreidler. 1979. Creating new words by shortening. *English Linguistics*, 13:24–36.

Rich Ling and Naomi S. Baron. 2007. Text messaging and IM: Linguistic comparison of American college data. *Journal of Language and Social Psychology*, 26:291–98.

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

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15:287–333.

Crispin Thurlow. 2003. Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1).

Kristina Toutanova and Robert C. Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proc. of ACL 2002*, pages 144–151. Philadelphia.