Beyond Normalization: Pragmatics of Word Form in Text Messages

Tyler Baldwin
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
baldwi96@cse.msu.edu

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

Abstract

Non-standard spellings in text messages often convey extra pragmatic information not found in the standard word form. However, text message normalization systems that transform non-standard text message spellings to standard form tend to ignore this information. To address this problem, this paper examines the types of extra pragmatic information that are conveyed by non-standard word forms. Empirical analysis of our data shows that 40% of non-standard word forms contain emotional information not found in the standard form, and 38% contain additional emphasis. This extra information can be important to downstream applications such as text-to-speech synthesis. We further investigated the automatic detection of non-standard forms that display additional information. Our empirical results show that character level features can provide important cues for such detection.

1 Introduction

Text message conversations are often filled with non-standard word spellings. While some of these are unintentional misspellings, many of them are purposely produced. One commonly acknowledged reason that text message authors intentionally use non-standard word forms is to reduce the amount of time it takes to type the message, or the amount of space the message occupies.

This phenomenon has motivated the text message normalization task (Aw et al., 2006), which attempts to replace non-standard spelling and symbols by their standard forms. The normalization task is potentially critical for applications involving text messages, such as text-to-speech synthesis.

However, one important aspect that is overlooked when performing normalization is the use of non-standard word forms to express additional information such as emotion or emphasis. For instance, consider the following text message conversation:

A: They won the game!

B: Yesssss

The intent of the utterance by person B seems clear: he wishes to show that he is happy about the event described by person A. If the non-standard form *Yesssss* was normalized to the standard form *yes*, the intent conveyed by the utterance would be ambiguous; it could suggest that person B is happy about this turn of events, or he is indifferent, or he could simply be acknowledging that he already knows this fact. By using the non-standard form instead of the standard one, Person B communicated his excitement to A.

As shown in the above example, text message users often employ these non-standard forms to display extra pragmatic information that is not easily displayed otherwise. However, because normalization is only concerned about converting non-standard spellings to standard forms, it has the potential to remove this important pragmatic information.

To address this problem and to better understand some of the pragmatics of non-standard spellings in text messages, we conducted an initial investigation. In this study, we investigate the prevalence of non-standard spelling for the purpose of displaying information not captured in the standard word form. We also investigate the non-standard word form style associated with extra information and make a first attempt at identifying whether a non-standard form holds extra pragmatic information.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1437–1441,
Chiang Mai, Thailand, November 8 – 13, 2011.

2 Related Work

There are two main areas of related work: text normalization and affective text classification. Because it may be unclear how non-standard forms should be read aloud, the field of text-to-speech synthesis has long been interested in text normalization. Sproat et. al. (2001) study several different corpora and identify several types of non-standard word, including several seen frequently in text message data, such as misspelling, abbreviation, and "funny spellings". More recent work (Zhu et al., 2007) has employed conditional random fields in an attempt to handle word normalization simultaneously with several related problems such as detecting sentence and paragraph boundaries.

Several different approaches have been proposed for normalization of text messages specifically, including those motivated by machine translation (Aw et al., 2006) and spell-checking (Choudhury et al., 2007). Most recently, Pennell and Liu (2010) use handcrafted rules as classification features to normalize SMS terms that contain character deletion, with a focus on normalization for text-to-speech systems. A few hybrid approaches (Kobus et al., 2008; Beaufort et al., 2010) and an unsupervised approach (Cook and Stevenson, 2009) have also been investigated. All of these methods assume that the normalized form is functionally equivalent to the non-standard form found in the text; none address the potential existence of extra information in the non-standard form.

Affective text classification attempts to identify the type or polarity of emotion that is expressed by the text, without the aid of extra linguistic cues such as gesture or prosody. Kao et. al. (2009) survey the field and divide approaches into 3 categories: 1) keyword based approaches (Bracewell, 2008), 2) learning-based approaches (Alm et al., 2005; Yang et al., 2007; Binali et al., 2010), and 3) hybrid approaches (Wu et al., 2006; Agarwal et al., 2009). Although there has been some recognition of the effect that non-standard word forms play in emotion detection (Zhang et al., 2006), the primary feature sources for emotion detection systems has been at the word and sentence level (Quan and Ren, 2010). To our knowledge, no previous work has focused on the role non-standard word form plays in conveying emotional and other pragmatic information in text messages.

Figure 1: Example dialogue from our corpus

3 Empirical Analysis

3.1 Data Set

In order to access whether non-standard word forms have additional pragmatic information, it is necessary to study these forms in their original dialogue context. Because no currently available text message dataset contains messages in context, we collected our own. The website "Damn You Autocorrect"[1] posts screenshots of short text message conversations that contain mistakes produced by automatic spelling correction systems. To create an initial dataset, 1190 text message conversations were transcribed. A sample dialogue is shown in Figure 1.

The speech bubbles originating from the left of the image in Figure 1 are produced by one participant, while those originating from the right are produced by the other. The dialogue shown contains several examples of non-standard spelling. The non-standard form *lookin* drops the letter *g* from the end of the morpheme *ing*, a technique commonly used in informal writing. Two other non-standard spellings, *hiii* and *goooooood* exemplify the use of letter repetition. This dialogue also includes the common texting slang term *lol*.

Since we are interested in studying the presence of extra information in non-standard word forms, we must first identify word forms that contain non-standard spelling. To create a set of non-standard word forms, we used the CMU pronouncing dictionary[2] as a vocabulary set and selected all tokens that were out of our vocabulary. Those tokens that were simply legitimate words in the lexicon, such as proper names or obscure terms not in our dictionary, were manually removed. This left us with a data set of 764 non-standard word tokens.

[1] www.damnyouautocorrect.com

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

3.2 Survey

To assess which word forms displayed extra pragmatic information, we created a short survey that asked users of Amazon Mechanical Turk[3] to determine whether this form contained information not present in the standard form. Survey participants were given the word in context and asked to answer the following four questions:

1. "What is the standard form of this word?"
2. "What type of emotion, if any, is provided by the spelling used that is not provided by the standard form?" (Choose from the following: none, fear, suprise, happiness, disgust, sadness, anger, other)
3. "What other information, if any, is provided by the spelling used that is not provided by the standard form?" (Choose from the following: friendliness/closeness, emphasis, other, none)
4. "Why do you think the writer chose to use the modified spelling instead of the standard form?" (Choose from the following: wanted to display extra information, wanted to save time or space, made an unintentional mistake, other)

Three separate annotators were asked to examine each word form. The observed agreement between any two annotators was around 80% for a given question. For our analysis, we consider a case in which 2 or more annotators agreed as the gold standard. Cases in which no annotators agreed were thrown out, judged separately for each question[4].

3.3 Analysis Results

Figure 2 shows the results of question 2. The emotions used include the six basic emotions (Ekman, 1993) often used in affective text literature. If several emotions were displayed, annotators were asked to pick the emotion that was displayed most strongly. As shown, 5 of the 6 emotions were present in our corpus, with only fear being absent. Although many forms did not contain extra emotion, a full 40% of them did. When additional emotional information was present, it was most commonly positive; happiness was by far the most common emotion displayed.

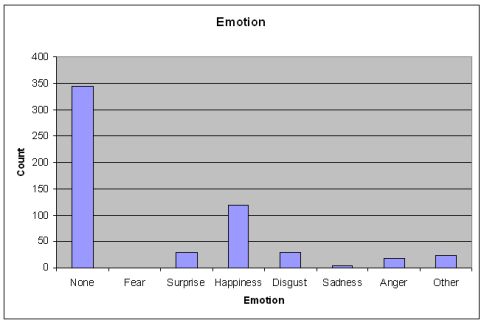

Figure 2: Distribution of forms containing emotion not present in normalized form

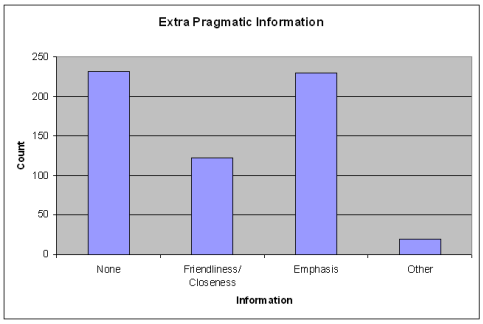

Figure 3: Distribution of forms containing additional information not present in normalized form

Figure 3 shows the results of question 3. Although it was again common for no extra information to be present, cases in which non-standard forms were used to emphasize a word were nearly as common, appearing in 38% of our instances. The use of non-standard forms to express emphasis appears to be widespread in text messaging data. This is an important finding, especially relevant to text-to-speech research. Additionally, Figure 3 suggests that another common usage of non-standard forms, found in 20% of our data, is to display a sense of kinship with the reader through subtle expressions of friendliness or closeness.

Results for question 4 are shown in Figure 4. Wanting to display extra information was perceived as a primary reason why text message authors chose a non-standard spelling. This seems to suggest that, in choosing non-standard word forms, expressiveness is a primary concern for text message writers.

Overall, the results in the figures suggest that the need for greater expressiveness is a paramount reason why text message writers choose non-

[3] mturk.amazon.com

[4] This accounts for the difference in total instances between Figures 2, 3, and 4

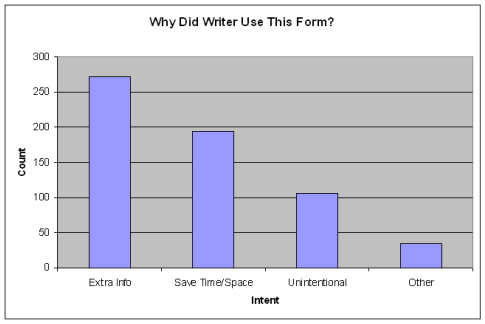

Figure 4: Perceived intent of text message author

standard spellings. It is thus relevant for text message normalization systems to consider the intent of the writer in producing a non-standard form, to ensure that the intended meaning is not lost. This leads to the question of whether an automated system can accurately recognize words that carry extra pragmatic information. In the next section, we take an initial look at this problem.

4 Automatic Identification of Words with Extra Information

We model the task of identifying whether a non-standard word form is intended to display extra information as a binary classification task. All instances that were marked by annotators as having some form of emotion or extra information were considered to be positive instances.

We drew features from three sources for our classification: character level features, punctuation features, and positional features. Because we are focused on classifying the emotional or pragmatic content of the word and not the utterance, we restrict our feature set to only features that pertain the to the word itself.

Character level features. Our feature set focused primarily on character level features. Several features focused on identifying the type of abbreviation used. Features indicating whether the word contained the same letter repeated more than twice, the maximum number of times a letter was repeated in the word, and whether deletion of repeated characters produced an in-vocabulary word were used to detect cases of word elongation. Edit distance from the closest word using only insertions was used as an indicator of word shortening and truncation. One additional feature recorded whether the non-standard form was longer than the normalized form. Features were also included to detect whether the non-standard form contained concatenated words or contained numbers or non-alphanumeric characters. Whether or not a word was written in all capital letters was also observed.

Punctuation features. Punctuation features capture some information beyond that of the word form. The punctuation features detected whether the word was followed by a comma, period, question mark, exclamation point, or emoticon.

Positional features. Positional features were the most discourse dependent features examined. These features indicated whether the word was the first, last, or only word in the current message.

Classification was performed using an SVM classifier. Ten-fold cross validation was performed. The results are shown in Table 1. A majority class baseline suggests that classification is not trivial; although many instances carry extra information, many do not. As shown, the use of character level feature alone achieves above baseline performance of 72.4% ($p < 0.01$). Adding additional features on top of this does not result in an increase in performance.

	Accuracy
Baseline	59.5%
Character Level Features Only	72.4%
Character Level + Punctuation	72.3%
All Features	72.4%

Table 1: Classification of word forms by the presence of added information

5 Conclusion

The analysis presented in this paper shows that non-standard word forms contain additional pragmatic information not present in the standard form. Some of the main functions of this extra information include the expression of emphasis, happiness, and friendliness. It is important that text message normalization systems recognize and address this fact, as it is relevant for downstream applications such as text-to-speech synthesis.

Additionally, this work introduced the problem of identifying whether a non-standard text messaging form was intended to display pragmatic information beyond that of the base form. Our initial investigation showed that above baseline performance could be achieved, but that the problem was non-trivial and required further study. Future work is needed to more robustly address this problem, as well as more closely examine the relationship between non-standard spellings and individual types of emotional and other pragmatic information.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32, Athens, Greece, March. Association for Computational Linguistics.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 579–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, Uppsala, Sweden, July. Association for Computational Linguistics.

H. Binali, Chen Wu, and V. Potdar. 2010. Computational approaches for emotion detection in text. In *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*, pages 172 – 177.

D.B. Bracewell. 2008. Semi-automatic creation of an emotion dictionary using wordnet and its evaluation. In *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*, pages 1385 –1389.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *Int. J. Doc. Anal. Recognit.*, 10(3):157–174.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78, Boulder, Colorado, June. Association for Computational Linguistics.

Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48:384–392.

E.C.-C. Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. 2009. Towards text-based emotion detection a survey and possible improvements. In *Information Management and Engineering, 2009. ICIME '09. International Conference on*, pages 70 –74.

Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one ? In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 441–448, Manchester, UK, August. Coling 2008 Organizing Committee.

D.L. Pennell and Yang Liu. 2010. Normalization of text messages for text-to-speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4842 –4845.

Changqin Quan and Fuji Ren. 2010. An exploration of features for recognizing word emotion. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 922–930, Beijing, China, August. Coling 2010 Organizing Committee.

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

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. 5:165–183, June.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 275–278, Washington, DC, USA. IEEE Computer Society.

Li Zhang, John A. Barnden, Robert J. Hendley, and Alan M. Wallington. 2006. Developments in affect detection in e-drama. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 203–206, Stroudsburg, PA, USA. Association for Computational Linguistics.

Conghui Zhu, Jie Tang, Hang Li, Hwee Tou Ng, and Tiejun Zhao. 2007. A unified tagging approach to text normalization. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 688–695, Prague, Czech Republic, June. Association for Computational Linguistics.