Cognitive Computation (2020) 12:778–792
https://doi.org/10.1007/s12559-020-09723-7

REVIEW

A Review of Shorthand Systems: From Brachygraphy to Microtext and Beyond

Ranjan Satapathy[1] · Erik Cambria[1] · Andrea Nanetti[2] · Amir Hussain[3]

Received: 10 September 2019 / Accepted: 5 March 2020 / Published online: 22 June 2020

Abstract

Human civilizations have performed the art of writing across continents and over different time periods. In order to speed up the writing process, the art of shorthand (brachygraphy) came into existence. Today, the performance of writing does not make an exception in social media platforms. Brachygraphy started to re-emerge in the early 2000s in the form of microtext in order to facilitate faster typing without compromising semantic clarity. This paper focuses on microtext approaches predominantly found in social media and explains the relevance of microtext normalization for natural language processing tasks in English. The review introduces brachygraphy and how it has evolved into microtext in today's social media–dominant society. The study provides a comprehensive classification of microtext normalization based on different approaches. We propose to classify microtext based on different normalization techniques, i.e. syntax-based (syntactic), probability-based (probabilistic) and phonetic-based approaches and review application areas, strategies and challenges of microtext normalization. The review shows that there is a compelling similarity between brachygraphy and microtext even though they started centuries apart. This paper represents the first attempt to connect brachygraphy to current texting language and to show its impact in social media. This paper classifies microtext normalization according to different approaches and discusses how, in the future, microtext will likely comprise both words and images together. This will expand the horizon of human creative power. We conclude the review with some considerations on future directions.

Keywords Microtext normalization · Shorthand systems · Brachygraphy · Social media processing

Introduction

Brachygraphy is a system of writing using abbreviations or special characters. Writing can be construed as a human art or craft ($\tau\acute{\epsilon}\chi\nu\eta$/technē) that in historical times involved skill in marking letters, words, ideograms or symbols with different instruments depending on the writing materials derived from the animal parts, minerals and vegetal[1] that were used by the scribes. Writing tools, materials and purposes apart, a variety of abbreviation systems exist attested to different writing systems due to both speed up the writing process and save valuable writing surface. For example, Ancient Greek and Latin scripts developed a series of abbreviation methods that were later shared by Romance languages[2] and Germanic languages[3].

The work related to brachygraphy is also highlighted in Peter Bales's study [31], published in three editions. The first two editions published in 1590 and 1597 respectively provide stenographic methods, and the third edition published in 1600 adds abbreviation tools [31]. The abbreviations have been adopted into today's English

✉ Erik Cambria
cambria@ntu.edu.sg

Ranjan Satapathy
satapathy.ranjan@ntu.edu.sg

Andrea Nanetti
andrea.nanetti@ntu.edu.sg

Amir Hussain
A.Hussain@napier.ac.uk

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[2] School of Art, Design and Media, School of Humanities, Nanyang Technological University, Singapore, Singapore

[3] School of Computing, Edinburgh Napier University, Edinburgh, UK

[1] https://en.wiktionary.org/wiki/vegetal

[2] https://en.wikipedia.org/wiki/Romance_languages

[3] It includes English, before and after the advent of print

language[4] also. A few ligatures are even more common, globally. For example, the English ampersand (&) is an ancient Latin ligature of the cursive letters "e" and "t" forming the Latin word *et* (and) as shown in Fig. 1. The symbol @ used in e-mail accounts between the user's name and the domain name is an ancient Latin ligature of the cursive letters "a" and "t" forming the Latin word *at* (but) shown in [12].

The brachygraphy used as a means to communicate is now termed as microtext. The transition of writing occurred during the so-called Industrial Revolution affected both longhand and shorthand technologies because of the creation and commercialization of handy devices that mechanized human writing methods for business and government services. The technological discoveries in this field have long roots and are the result of extensive and still ongoing implementation processes. The Wikipedia pages on "typewriter" and "stenotype" perfectly showcase and inform better than any other paper or book made accessible in English by academic, commercial publishers. In 1874 [17], "Sholes & Glidden Type-Writer" were successfully commercialize, and in 1879[5] patented his "Stenograph Shorthand Machine". Wikipedia informs that a first shorthand machine was invented in Germany by Karl Drais in 1830, and several pioneering devices were created and patented in Italy (by Antonio Zucco in 1863), USA (by Miles M. Bartholomew in 1879), France (by Marc Grandjean in 1909), until 1913, when Ward Stone Ireland made the stenotype, which is considered the ancestor of all modern shorthand machines. Table 1 compares the taxonomy used in medieval and renaissance Latin and today's social media platforms.

In Table 1, it can be seen that most of the shortened format of words in Medieval and Renaissance Latin are still in use but has variations according to geographic location.

The impact of social media and SMS is increasing in our daily lives. These sources provide the analysts with a large amount of text data for data mining and machine learning. However, this data is notoriously noisy as people use a lot of shorthand language and hence destroying its utility for analyzing. Hence, it is essential to convert this noisy text into their respective standard text. Today, shorthand writing is commonly referred to as "microtext", i.e. a branch of natural language processing (NLP) that focuses on handling "semi-structured" texts. The tasks discovered in this branch overlap considerably with those in more traditional NLP [11], including topic detection, summarization, sentiment analysis and classification, question-answering, and

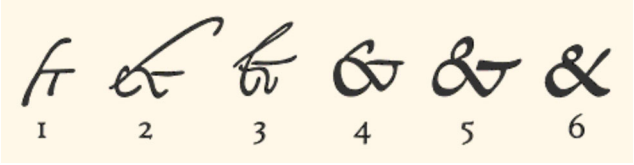

Fig. 1 The ligature "et" in the shorthand system developed by Tiro to annotate M. T. Cicero's speeches (Rome, First Century BCE). Image source: [54]

information extraction [22]. Microtext in social media is a form of contemporary brachygraphy. The term *microtext* was introduced by US Navy researchers [69] to define a type of written text that has three main characteristics:

(a) It is compact, typically one or two sentences, and possibly as little as a single word (abbreviations like "hru" for "how are you");
(b) It is written informally and thus may use relaxed grammar, a conversational tone, vocabulary errors and uncommon abbreviations and acronyms; and
(c) It is semi-structured, which means it might include metadata like a period or author information.

Analysis of microtext falls under an interdisciplinary research area, as shown in Fig. 2, viz., it utilizes knowledge from different domains. Table 2 depicts the common trends used to generate microtext. Given that most data today is mined from the Web, and the text classifiers are trained in everyday English microtext analysis is a key for many NLP and data mining tasks. Microtext has become omnipresent in today's world, some of the sources are listed below:

(a) Short Message Service (SMS)
(b) Instant Messaging such as Microsoft Messenger
(c) Multi-User Chatrooms
(d) Voicemail Transcriptions such as Google Voice
(e) Microblogs such as Twitter, Weibo, Facebook and Google+.

Since a subset of microtext is heavily based on phonetics [70], it is largely colloquial-dependent. The same set of characters could have a completely different meaning in different languages, e.g. "555" is negative in Chinese language (because the number "5" is pronounced as "wu" and "wuwuwu" resembles a crying sound) but positive in Thai language (as the number 5 is pronounced as "ha", and three consecutive 5s correspond to the expression "hahaha"). Table 3 depicts some real-time sample messages.

The work of [27] is a well-known reference to the task of lexical normalization of tweets, albeit their study focused on English tweets for one-to-one normalization.

In this context, we propose to compare the taxonomy of the Latin system of abbreviations used in Medieval and

[4] www.en.wikipedia.org/wiki/List_of_Latin_abbreviations (accessed on 15 July 2019)

[5] http://americanhistory.si.edu/collections/search/object/nmah_849951 (accessed on 15 July 2019)

Table 1 Comparison of the taxonomy between Latin text and today's social media

Medieval and Renaissance Latin	Today's social media
Abbreviations by truncation ("s.p.d." salutem plurimam dicit)	Acronyms ("NTU" Nanyang Technological University)
Abbreviations by contraction ("pbr"presbyter)	Clipping ("apl" apple)
Abbreviation symbols with proper meaning (the sign with a shape of "9" means con and the one with the shape of "7" means	–
Abbreviation symbols with variable meaning (the sign with a shape of "3" at the end of a word can mean m, et, is according to the context)	Phonetic Substitution (the numeral "2" means "to" in "2gether" and "too" in "me 2")
Abbreviations by superscript letter (the symbol "a^i" for aliqui)	–
Conventional symbols ("÷" means est and "=" means esse)	–
Symbols for numerals, Roman (I, II, III, etc.) and Arab (1, 2, 3, etc.)	Already incorporated in English

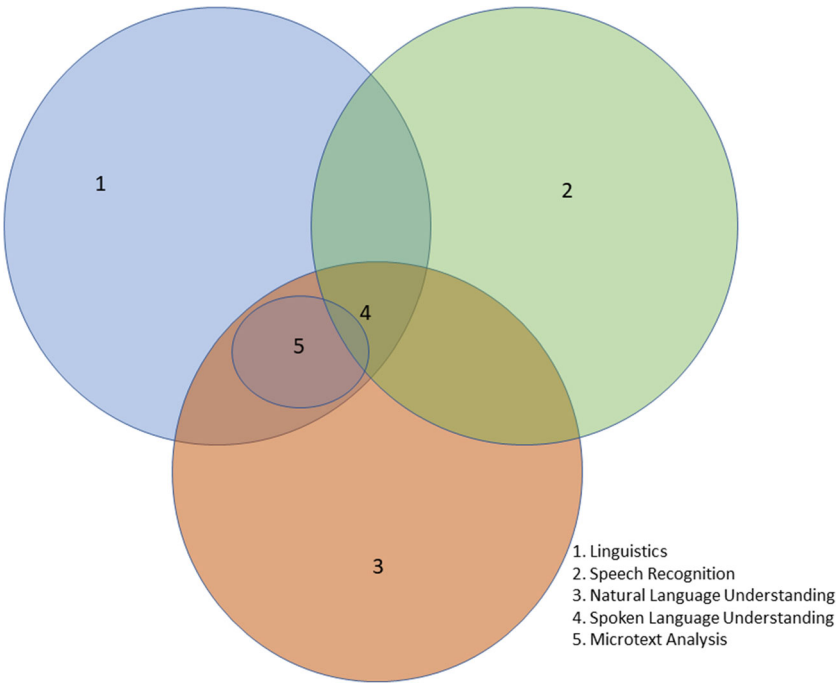

Fig. 2 Microtext analysis as interdisciplinary approach

Table 2 Common trends in microtext

Spelling	Ungrammar	Non-standard words
Typo error	Substitution of shorter words	Use of emotions
Deletion of whitespace between two words.	Deletion of pronouns (especially subject)	Abbreviated words
Use of number in words	–	Features from spoken languages

Table 3 Real-time sample messages depicting the use of microtext

I'll meet u b4 lec then...
Where r u
Hey r v goin out tmr
So u stayin in d hstl ?
R u going to b done anytime soon ?

Renaissance Europe [73] with the microtexts used today in social media as shown in Table 1.

Overall, the specific contributions of this review are as follows:

(a) We systematically compare several approaches to solve the microtext normalization problem. We also cautiously provide a clear broad definition for brachygraphy and how it is similar to microtext.
(b) To our best knowledge, this review provides the most comprehensive list of fundamental theories that can be utilized when studying microtext normalization.
(c) This review comprehensively and extensively studies microtext analysis presenting (i) approaches to qualitatively and quantitatively analyze, detect and normalize the microtext; (ii) datasets; and (iii) application of microtext normalization, followed by challenges and what's beyond it.

Approaches to Microtext Normalization

We have classified microtext is divided into syntax-based (syntactic), probability-based (probabilistic) and phonetic-based approaches (Fig. 3).

Syntax-Based Approaches

There are several elements in text that can have syntactic value. It can be characters, words, concepts or phrases. This subsection introduces to the syntax-based approaches in microtext normalization. The overview of the syntactic-based approaches is shown in Table 4.

Authors in [20] find most probable candidate from database for a particular OOV after applying Levenshtein distance. For example, if a word replacement from the database is available, it becomes the most probable solution for that NIV. In addition to this, authors also incorporated rules for the elongated texts like "goooooooood" instead of "good".

The authors in [26] investigated properties of lexical transformations as observed within the context of microtext used in Twitter. The data included were scraped from Twitter API. They followed a two-pronged analytic approach:

1. Firstly, they conducted a context-free linguistic analysis of all words not usually found in everyday English language.
2. Secondly, authors conduct a more in-depth and contextual word-level analysis by first normalizing the noisy message to recover the most probable standard form of the message and noting down the varieties of changes that were produced, and then examining these changes across different general contextual dimensions based on client and geographic location.

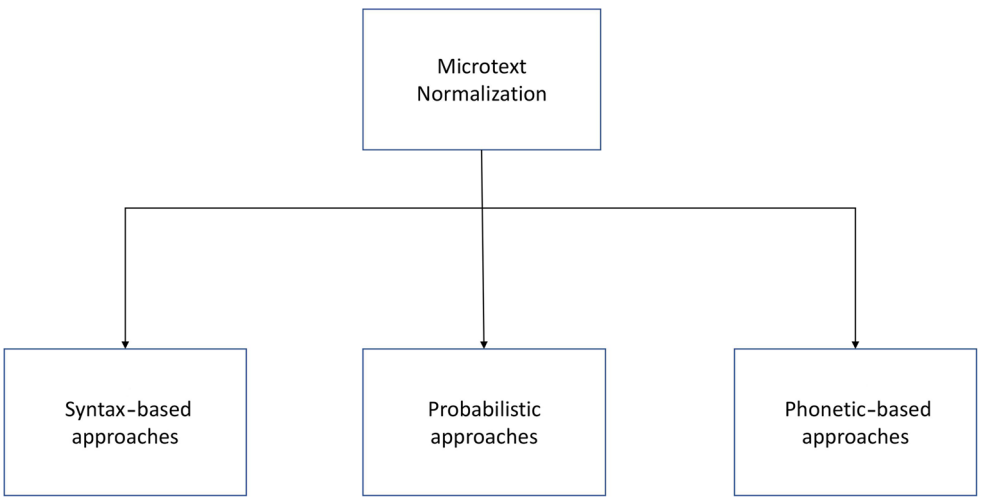

Fig. 3 Outline of microtext normalization tasks

Table 4 Research overview of syntax-based approaches to microtext normalization

Algorithm	Dataset	Evaluation metric	Algorithm rule types
[20]	SMS and Twitter	Levenshtein distance	Letter shortening + database + lexical matching
[26]	Twitter	Lexical transformation	Rule based + lexicon
[58]	English Gigaword corpus + CMU lexicon	Accuracy	Rule based+Maxent classifier+lexicon
[53]	SMS	Top-n	Abbreviation lookup + phonetic algorithm
[16]	Twitter	Average errors	Rule based
[70]	Twitter and SMS	Accuracy	Rule + lexicon+soundex
[83]	Twitter, SMS and call-centre data	Precision and Recall	Generator based algorithm
[9]	Twitter	Inter-rater agreement	Graph+lengthening
[63]	North Sàmi and Greenlandic Wikipedia	Coverage + word error rate	Finite-state model with weights
[36]	Twitter and SMS	Sentence level Accuracy	Database and a user feedback system
[6]	SMS data	Word Error Rate and BLEU score	Morphological analysis+ contextual disambiguation
[80]	Twitter and SMS data	F-Measure and Accuracy	Orthographic factor, a phonetic factor, a contextual factor and acronym expansion

The authors in [58] built a normalization system for text messages to be read by a text-to-speech (TTS) engine. They anonymously collected around 20000 sentences from volunteers utilizing a Telit GM862-GPS cell phone modem. The approach involves rules to translate English to texting language and automatic abbreviation generation using a maxent classifier.

The authors in [53] propose a three-level architecture to deal with microtext in QA systems, which is described below:

1. Authors remove noise present in the SMS tokens. They combined Soundex, Metaphone algorithm with a modified LCS algorithm;
2. Authors select a semantically similar set of candidate questions and;
3. They build a syntactic tree matching (STM) and WordNet-based similarity to improve the results.

In [16], authors evaluate the performance of two leading open-source spell checkers on data taken from the microblogging service Twitter and measure the extent to which their accuracy is improved by pre-processing with their system. The system "Casual English Conversion System" (CECS) is designed on the basis that errors and irregular language used in casual English found in social media can be grouped into several distinct categories mentioned in Section "Challenges in Microtext Normalization". As the language is dynamic, the usage of microtext has also changed a lot since [16]. Authors in [70] show an application of sentiment analysis with rules and lexicon in addition to the Soundex algorithm. The results show an increase in accuracy for sentiment analysis task by 4%. In [83], the authors aim at converting raw informal texts into their correct grammatical version using a parser-centric word-to-word normalizations method. The authors tie normalization performance directly to parser performance. The framework allows for transfer to new domains with a minimal amount of data and effort. The dataset includes Twitter, SMS and call-centre data. The evaluation is done with the word error rate and BLEU score.

Authors in [9] introduce an automated method which correlates word lengthening to subjectivity and sentiment. The lexicon is built mainly for Twitter and related social messaging networks. They present a method to identify domain-specific sentiment-bearing and emotion-bearing words. The evaluation is done with human annotation. This paper misses out other aspects, like the link between the length and the strength of individual words.

In [63], the authors propose a finite-state spell-checking as an alternative to the conventional string-based algorithms. North Sámi, Finnish, English and Greenlandic languages are included in this study. For the English language model, authors use the data from [56] and [62], which is a basic language model based on a frequency weighted wordlist obtained from freely accessible Internet corpora such as Wikipedia[6] and project Gutenberg[7]. The language models for North Sámi, Finnish and Greenlandic are drawn from the open-source repository[8] of finite-state language models. Improvements to Mays, Damerau and Mercer (MDM) model was addressed in [78] by incorporating a trigram-based real-word SCC method of MDM. The proposed method performs better than the WordNet-based method of [29].

[6] http://en.wikipedia.org

[7] The Project Gutenberg website http://www.gutenberg.org/

[8] http://giellatekno.uit.no

The framework in [6] is based on morphological analysis and contextual disambiguation developed in the context of an SMS-to-speech synthesis system. The evaluation was performed on the corpus of 30,000 French SMS by ten-fold cross-validation [41].

The authors [36] introduced a lexico-syntactic normalization model for cleaning the noisy texts. The channelized database and a user feedback system are utilized for the normalized system. The syntactic analysis of sentences is based on a bottom-up parser. The model captures the user interaction for improving the model accuracy. The noisy input sentence is tokenized and parsed through lexicons to transform it into the standard token, and then the output of the lexicon is combined to form the IV sentence. The lexical approaches as such are critical because they need to contain different forms of each OOV words which tend to evolve rapidly. The lexicon is a difficult task to maintain with the increasing usage of social media languages.

The authors [80] propose a method which combines four factors, namely, an orthographic factor, a phonetic factor, a contextual factor and acronym expansion. Authors formulate each factor as channels. The model is the combination of all the four channels with two variations, namely:

1. Generic Channel, and
2. Term Dependent Channel.

The authors evaluated the proposed model on Tweets and SMS messages at the word level.

Probability-Based Approach

The last few years have seen a significant shift in speech and language technology as it is being taken over by deep learning approaches. The overview of probability-based approaches is shown in Table 5.

The authors [43] propose a two-stage translation method from OOV to IV at word level. This framework leverages on phonetic information, where non-standard words translated to possible pronunciations, which are then mapped to standard words. Their results show that this approach enhances the system's resistance to detect and transform OOV words into their IV form. Authors show their results on SMS and Twitter messages. The authors in [15] propose a supervised text normalization model based on learning edit operations which incorporates features from unlabeled data via character-level neural text embeddings. The text embeddings are generated using a Simple Recurrent Network, and it claims to improve on state-of-the-art with less training data and with no lexical resources. The novelty of the paper lies in utilizing less training data and no lexical resource with simple RNN. In [34], the authors use HMM to convert a letter-to-phoneme features. The authors used CMUdict, Celex (English, Dutch and German) and Brulex (French). The experiments were conducted with instance-based classifier [2] for predicting phonemes implemented in the TiMBL package [19]. They utilized the HMM technique for post-processing instance-based learning to provide a sequence prediction. In [60], the authors introduce a two-phase approach for normalization of abbreviations in informal texts. The first phase uses an abbreviation model to generate possible candidates, and the second stage uses a language model to choose the best candidate. The authors in [42] participated in W-NUT Lexical Normalization for English Tweets challenge[9] which combines two augmented feedforward neural networks, a flag that acts as an identifier for the words to be normalized and a normalizer, which takes a single token at a time and outputs a corrected version of that token. Authors show that their system achieved an F1-score of 81.49% trailing the second-place model by only 0.26%. The work done in [81] is a preliminary text normalization technique which uses pre-neural approaches. The authors characterized the relationship between standard and non-standard tokens by a log-linear model, permitting arbitrary features. Authors in [47] introduce a social media text normalization hybrid word-character attention-based encoder-decoder model. The proposed character-based component is trained on synthetic adversarial data that are designed to capture errors commonly found in the online user-generated text. The authors use two encoder-decoder models: a word-based Seq2Seq model and a char-based Seq2Seq model. First, the tokens are passed through the word-based Seq2Seq model, while for transforming tokens not found in the word-level model's vocabulary, it either backtracks to a secondary character-based Seq2Seq model if its confidence is high or copies the source token. Authors use W-NUT [3] dataset. The authors in [14] constructed an HMM for every word in the standard language, to represent all possible variations of the corresponding *texting language* according to their associated observed probabilities. The model uses a word-level decoder to transform OOV English SMS text to their standard English forms with an accuracy of 89%. The decoder is used for automatic correction as well as information extraction and retrieval from noisy English documents such as e-mails, blogs, wikis and chat logs. The structure of the HMM is based on linguistic analysis of the SMS data. The authors evaluated the word-level model on 1228 different tokens collected from the SMS corpus that does not exist in the training set. The probabilities of every test samples are computed using the Viterbi algorithm. In [82], the authors propose a hybrid method for multi-class sentiment analysis of micro-blogs, which combines the model and lexicon-based approach. The authors combined the effect of emoticons and Naïve Bayes classification to

[9] https://noisy-text.github.io/norm-shared-task.html

Table 5 Research overview of probability-based approaches to microtext normalization

Algorithm	Dataset	Evaluation metric	Algorithm rule types
[34]	Celex(English, Dutch, German) + Brulex (French) + CMUdict (English)	Word accuracy	Expectation-maximization + HMM
[15]	Twitter	Character level	Simple Recurrent Network
[42]	Twitter		Feed Forward Neural Networks
[60]	SMS and Twitter Data	Top-N	CRF + Language Model
[50]	WhatsApp messages	Word-level	Encoder-Decoder Model
[71]	Twitter dataset	Character level	Seq2Seq model
[4]	CELEX and NETtalk	Word accuracy (WA) and syllable break error rate (SBER)	Positional tags capture and Structural tags for Letter-To-Phoneme
[43]	SMS data and Twitter	Top-n	character-based machine translation approach
[81]	Twitter data	Recall and precision	Monte Carlo training algorithm
[47]	LexNorm dataset	Precision, recall and F-score	Character and word-level seq2seq models
[14]	SMS data	Accuracy	Lexicon with Hidden Markov Model
[82]	Chinese microblogging site Sina	Precision and recall	Naïve Bayes classifier and a lexicon-based classifier
[27]	SMS corpus and Twitter	F-score	Lexicon + word similarity + context support
[24]	Pair of misspelled words+ WSJ Penn Treebank	Coverage, F-score	Hidden Markov Model
[48]	WMT'15	BLEU score	Word + character-level NMT models
[59]	SMS data	Top-N	Character-level MT + a language model
[79]	LexNorm1.1 and LexNorm1.2	precision, recall and F-score	Lexicon + Hidden Markov Model
[39]	University of Aix-en-Province [30] + Catholic University of Louvain [23]	Word error rate	Machine Translation + ASR technique
[4]	CELEX and NETtalk	Word accuracy and syllable break error rate	SVM Hidden Markov Model
[37]	Twitter data	BLEU and NIST score	Orthographic Normalization+Syntactic Disambiguation+Machine Translation
[77]	Chinese-English MT and NUS SMS data	BLEU score	Hidden Markov Model + beam-search decoder

divide micro-blogs into three sentiments—positive, negative and neutral. The authors further divide negative opinion into: angry, sad, disgusted and anxious using sentiment dictionaries. They evaluate their algorithm on a Chinese microblogging site called Sina.

The authors in [27] proposed a classifier-based approach to detect ill-formed words and generate candidate words based on morphophonemic similarity. The most probable candidate for the word is chosen based on word similarity and context. The proposed method doesn't require any annotations and achieves state-of-the-art performance for an SMS corpus and a novel dataset based on Twitter. The authors do not normalize the abbreviations to their standard form. The authors trained an HMM in [24] with a mixed trigram model, where each state of the HMM is labeled either with a pair of part of speech (POS) tags or with a pair made up of a POS tag and a valid dictionary word. The output symbols from the HMM's states are the words observed in the input sentences. The valid words that label some of the states represent words in the confusion set of the observed words, collection of misspelt pairs and a corpus of correct sentences annotated for POS by both WSJ Penn Treebank and Stanford POS tagger [51]. Precision, recall and F-score evaluate the system. The training corpus also contains the following additional texts:

1. The New American Bible[10],
2. The Project Gutenberg Encyclopedia, and;
3. Four novels by Mark Twain[11].

[10]The US Conference of Catholic Bishops website: http://www.usccb.org

[11]The Project Gutenberg website: http://www.gutenberg.org/

Authors in [48] propose a word-character-based hybrid Neural Machine translation (NMT) which handles rare word translation as well. The authors compare purely word-based, purely character-based and hybrid models in which hybrid model scores the highest BLEU and performs better on rare word dataset [49]. The system achieves a new state-of-the-art result with 20.7 BLEU score on English-Czech translation task. An encoder-decoder model [50] to normalize Swiss German WhatsApp messages has also been proposed recently. A multilingual framework is still limited by dataset availability but has a lot of scope for the future. In a similar approach, recently authors [71] proposed a seq2seq-based encoder-decoder normalization framework in English. They show the improvement in sentiment analysis model by transforming OOV texts into their IV texts. The improvements in the sentiment analysis task show a requirement of the microtext module for the NLP tasks. Such evaluations are needed for other NLP tasks as well. The work done in [59] describes an SMT-based system for expanding abbreviations found in the informal text. The SMT follows a two-phase system. The first phase is trained at the character level, while the second phase is trained with an in-domain language model. In this way, the system learns mappings between character-level "phrases" and is much more robust to new abbreviations than a word-level system. In [79], the authors propose a syllable-based method for tweet normalization to study the cognitive process of non-standard word generation in social media. The authors segment the non-standard words c_i into syllables $s^1_{c_i} \cdots s^k_{c_i}$, and for standard syllable $s^j_{w_i}$ mapping to non-standard syllable $s^j_{w_i}$. The authors calculate the similarity by combining orthographical and phonetic measures. It combines the HMM channel model with four additional characteristics, namely, combination, syllable level, a priori knowledge and general patterns. The paper's main contribution is that the proposed normalization system relies on unlabeled samples, thereby making it much easier to adapt the method to handle non-standard words in any period of history.

In [39], the authors present a comparative study of methods aiming at normalizing the orthography of French SMS messages. Authors combine MT system with the ASR system to achieve 11% Word Error Rate on a test set of about 3000 unseen messages. The experiments reported by authors use two corpora. The first one has been collected at the University of Aix-en-Provence [30]; it contains approximately 9700 messages. The second corpus contains about 30000 messages [23] and gathered in Belgium by the Catholic University of Louvain. Both corpora contain the message and a reference normalization which has been produced and validated by human annotators. In [4], the authors trained support vector machine (SVM) for English syllabification. The proposed method improves the accuracy of the letter-to-phoneme conversion. Authors have employed two different approaches to tagging in this paper: one being positional tags [8] which captures where a letter occurs within a syllable; and second being structural tags [18, 74] expresses the role, each letter is playing within the syllable. Word error rate is shown to be reduced by 33% as well. The results reported are based on CELEX and NETtalk dataset. NETtalk and CELEX do not provide the same syllabification for every word. There are numerous instances where the two datasets differ in a perfectly reasonable manner (e.g. forging in NETtalk vs. forging in CELEX). In [37], the authors propose a syntactic normalization of Twitter Messages. They fed the pre-processed tweets into an SMT model to transform them into standard English. The tool that was used to build this system is Moses [40]. Moses is an SMT package which can produce high-quality translations from one language into another. The authors [77] focus on missing word recall and punctuation amendment. Authors propose a novel beam-search-based decoder for social media text normalization for SMT. The decoder effectively integrates different normalization operations. Authors have created two corpora: a corpus containing 1000 Weibo[12] messages with their normalizations in Chinese and their English translations; and another corpus which contains 2000 English SMS messages (NUS SMS corpus[13]) in [32].

Phonetic-Based Approaches

The biggest hurdle comes from texts that are affected by texting phenomena such as character repetition (for instance, "hiiii" for "hi") or phonetic-based character substitution (for instance, "dawg" for "dog"), to name just a few [76] (Table 6). Authors in [38] propose a phonetic tree-based microtext normalization on English Wiktionary. The proposed algorithm determines the probable pronunciation of English words based on their spelling. Thus, when the system encounters an out-of-vocabulary (OOV) word, it will determine the most probable in-vocabulary (IV) words with similar pronunciation. In [70], the authors demonstrated a phonetic-based algorithm to normalize tweets. They show that there is a high (>0.8) similarity index between tweets normalized by their proposed model and tweets normalized by human annotators, in 85.31% of cases. The system enhances the accuracy of the polarity detection module by $>4\%$.

In [33], authors propose a new word-searching strategy based on the idea of sounding out the consonants of

[12] A Chinese version of Twitter at www.weibo.com

[13] Available at www.comp.nus.edu.sg/~nlp/corpora.html

the word. The suggested algorithm uses a spelling and phonetic strategy to extract the base consonant data from both miswritten and real phrases. To serve as a reference, the first phase of their methodology extracts the VS and PS from valid English words. The algorithm also extracts streamlined word reconstructions with vowels re-inserted at the right locations. The second main phase of the methodology is to generate signatures of a new OOV phrase for which data about IPA is unknown. First, the writers accounted for three of the five original word transformations kinds. The third and final step of the methodology is to determine the respective IV words for the OOV phrases. They discover a collection of IV phrases with the same signatures together with their probability of occurrence. The algorithm then applies a number of heuristics in order to score the IV as a match to the OOV word.

In [72], authors propose a cognitive (phonetic) approach to solve the microtext normalization technique. They transform the concepts to their phonemic subspace by using a grapheme to phoneme converter. The proposed framework improves the accuracy of polarity detection by 6% as compared with the earlier model. In [44], the authors propose a cognitively inspired normalization technique that integrates different human aspects to normalize the OOV tokens. The method involves improved letter conversion, visual priming, and the resemblance between string and phone. The authors evaluated the system on both words- and message-level using four SMS and Twitter data sets. They also reveal that their method scores more than 90% word-coverage over all the four datasets and the comprehensive word-coverage can be successfully transposed into message-level performance gain.

Datasets

This section introduces to different datasets available for microtext normalization at different levels, namely, text level and phonetic level.

Text-Level Normalization

This subsection discusses different corpus created for text-level microtext normalization task:

(a) The corpus in [32] is created from SMS messages from three different sources. The messages are first obtained from a pool of 20 selected regular phone users for the corpus to have adequate depth per user. The age range of these users fall between 18 and 22, and has a collection of 6,167 messages altogether, which is about 60% of the messages in the corpus. Another group of messages is from the Yahoo SMS chat website, which constitutes about 602 messages, which shows the live SMS chat transcripts. The final group of messages was gathered from undergraduate students.

(b) There are 30000 SMS in the corpus [23] which is in the French language. These SMSs have been manually translated to create a bilingual corpus in which each message and its translated version is aligned.

(c) The SMS texts in [14] contain manually translated English forms and their non-standard English forms. The proper nouns (such as names of people, places) are replaced by a tag $< NAME >$ in both the non-standard language and translated (standard) language. The corpus text contains around 20000 tokens (words) out of which only 2000 are distinct. There are 83 characters in the SMS (non-standard) text for every 100 characters in the corresponding standard English text.

(d) The Edinburgh Twitter Corpus in [61] contains a large number of non-standard text but does not have corresponding standard English transcription.

(d) The small Twitter corpus used in [27] has also been released; this corpus has annotation and context but only contains 549 messages.

(e) Normalized Tweets constructed in the work [70] also follows word-level normalization. It contains 4000 non-standard and their standard text in the corpus.

Table 6 Research overview of phonetic-based approaches to microtext normalization

Algorithm	Dataset	Evaluation metric	Algorithm rule type
[38]	Twitter	Top-N	Phonetic tree-based framework
[70]	Twitter and SMS	Accuracy	Soundex-based framework
[33]	Wiktionary and TheFreeDictionary	Word-searching strategy	IPA based
[72]	Twitter	Accuracy	PhonSenticNet resource
[44]	SMS and Twitter data	Word- and message-level accuracy	Spell Checker + Character-level HMM

Phonetic-Level Normalization

The authors in [21] have done an extensive comparison study on different phonetic algorithms for the English language. In this subsection, we layout the different datasets available to do microtext normalization at the phonetic level:

(a) ARPABET is developed by Advanced Research Projects Agency (ARPA) as a part of their Speech Understanding Research project in the early 1970s. It is a mapping from IPA to "computer-friendly" ASCII symbols[14]. There are two representations in ARPABET: one adopts only one character and includes lower-case letters. The second uses only upper-case letters and is known as "2-characters".

(b) The TIMIT [46] corpus (originating in 1986) was collected to support the training and testing of ASR systems. The original distribution[15] is a diverse corpus of 630 American English speakers reading ten sentences each.

(c) The Carnegie Mellon University Pronouncing Dictionary (CMUdict)[16] is an open-source machine-readable pronunciation dictionary for North American English with over 134K words with their pronunciations. CMUdict is actively maintained and expanded regularly. Its entries are particularly useful for speech recognition and synthesis, as it has mappings from words to their pronunciations in the ARPAbet phoneme set. The current phoneme set contains 39 phonemes, and vowels carry a lexical stress marker:

0: No stress
1: Primary stress
2: Secondary stress

This phoneme set is based on the ARPAbet symbol set developed for speech recognition uses.

(d) A Lexicon built by [7] provides pronunciations for words and is called Festival. It consists of three distinct parts: an addendum, typically short consisting of hand added words; a compiled lexicon, typically large (10,000s of words) which sits on disk; and a method for dealing with words, not in either list (OOV words)[17].

[14] http://catalog.ldc.upenn.edu/docs/LDC93S1/PHONCODE.TXT
[15] http://catalog.ldc.upenn.edu/docs/LDC93S1/PHONCODE.TXT
[16] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[17] http://www.cstr.ed.ac.uk/projects/festival/manual/festival_13.html

Application of Microtext Normalization

This section dives into the areas of application where microtext normalization is used and implemented.

Information Retrieval

Information retrieval is the task of examining and retrieving desired knowledge from repositories of data. The first application of information retrieval is in search engines. Search engines aggregate information for a given query. The response to a query part is quite simple though underlying technologies and algorithms which make it possible are complex. However, a user is expected to query proper spelling which is not sustainable in the social media era. Hence, if the query itself is incorrectly spelt then a repository of five billion records and complex systems may not give a proper response for example if "Alan Turing" becomes "Alen Turnin" if "Captain America" becomes "Capt. Amerika" and if "together" is written as "tgthr". Review of query logs unveils plenty of misspelt words, wrongly split or merged words, dropping required quotations, unstemmed words, uncommon abbreviations without expansion and a lot of other colloquial variations. These are a serious problem since traditional search engines rely on context and string comparison to retrieve the matching list of information.

Text Classification

The text classification task is the class assignment of documents to a given set of classes. However, usage of microtexts distorts the content, and hence, the categorization performance gets affected. Some of the areas where text categorization is useful are call routing, categorization of hand-written client grievances and automated SMS routing.

Study in [1] shows the addition of the artificial noise and ASR noise into a standard text classification dataset had not much degradation in classifier performance. A generic system for text categorization was proposed based on a statistical analysis in [5]. They assessed their framework on the tasks of categorizing complaints from businesses and abstracts of paper-based German technical reports. Their method achieves 80% accuracy in classifying the texts and is very robust to typo.

Summarization

Selecting important sections or generation of new natural language from the given text comes under the umbrella of summarization of text. Selection of these methods is dependent on statistic, linguist and heuristic techniques.

Noisy text poses unprecedented hurdles to the summarization method. The difficulty of summarizing text documents that contain errors as a result of OCR has been studied in [35]. The addition of noise degrades the summarization system. Thus, pre-processing before summarization to normalize the microtext is important.

Information Extraction

The purpose of information extraction is to obtain structured information automatically, categorizing them into contextually and semantically well-defined data. The extracted information could be domain-dependent or independent depending upon the application. One of the major tasks in this is named entity recognition and extraction. It involves the extraction of entity names such as "people and organizations, place names, and temporal expressions". In [75], authors show that noisy text impacts information extraction system's performance. Authors in [52] studied the performance of named entity extraction under a variety of scenarios involving both ASR and OCR output. Their system was trained on both clean and noisy text to handle noise. The performance degradation is linear as a function of word error rates.

Sentiment Analysis

In recent years, sentiment analysis has become increasingly popular for processing social media data on online communities, blogs, wikis, microblogging platforms and other online collaborative media [10]. Sentiment analysis is a branch of affective computing research [66] that aims to classify text into either positive or negative—but sometimes also neutral [13]. Most of the literature is on the English language, but recently an increasing number of works are tackling the multilingual issue [45], especially in booming online languages such as Chinese [57].

Microtext normalization is a key task for sentiment analysis [11] (Fig. 4). Authors in [25] show that *Detection of human stress and relaxation* is crucial for timely diagnosing of stress-related diseases. It performs partial microtext normalization like SCC for missing spelling or repetitive letters which enhances the system performance. Authors in [70] show that microtext normalization plays an important part in polarity detection for tweets.

Challenges in Microtext Normalization

In [26], authors show that users exhibit different amounts of shortened English terms and different shortening styles depending upon the geolocation and culture they are brought up. The English text is essentially shortened in the following ways:

1. Using a shorter word form with similar pronunciation (phonetic variation);
2. Abbreviating a word;
3. Using only a prefix of a formal word.
4. Informal punctuation conventions including omitted and misused punctuation;
5. Redundant interjections;
6. Quotation-related problems, viz., omitted quotation marks;
7. "be" omission;
8. Tokenization problems; and
9. Informally written time expressions.

The challenge is not limited to the abovementioned ones. It also arises due to the use of colloquial terms. Taking Singapore as an example, the sentence endings "lah", "leh" and "lor" are commonly heard in Singlish conversations [28]. In addition to Singlish, concepts like "agak agak", means "estimate" in Malay, and "kaypoh", which is Hokkien for "busybody" [64]. Use of multilingual terms potentially confuses the classifier as to know the exact context of the word/concept used. Hence, it is an essential first step to mitigate in NLP. This review tries to consolidate all the abovementioned challenges into different umbrellas and showcase different solutions proposed by the authors. As the microtext evolves with the language and geographic locations, these challenges will keep on coming but in unique forms.

Beyond Microtext

The electronic computer radically changed the ways our society and economy work [68] at all levels. In brachygraphic terms, it seems that nothing changed yet. This review provides an insight into the work done in microtext normalization and its application. However, we also need to address what the future holds for this research field. Shall we expect the end of brachygraphy when the keyboard will be completely replaced by vocal dictation? Or are we entering into a more sophisticated scenario, that seems to have been already forecasted by the emojis, which (re)discovered the link between orality (words) and visual knowledge aggregation of ideas or emotions (images)?

Indeed, humans use both words and images to share the creations of their imagination (thoughts), and we know from both the humanities and cognitive sciences that human memory and imagination are strictly interrelated. In human communication history, over time and across space, we can see different relationships between text and images, with a

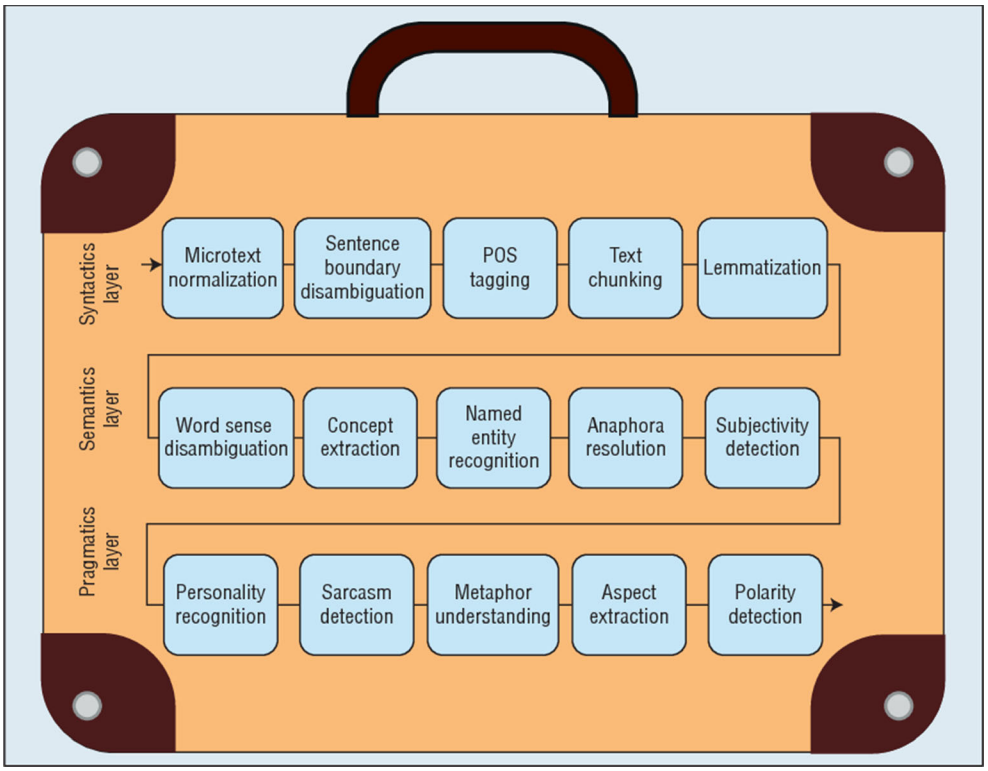

Fig. 4 Visualizing tasks of sentiment analysis as a suitcase [11]

predominance of one of them over the other and rarely a real emergence. This has already been a subject of discussion in pedagogy since Quintilianus' *De Institutio Oratoria* (XI, 2, 11–17) and Martianus Capella's *De Nuptiis Philologiae et Mercurii* (V, 538–539), where they refer to the famous story of Simonides of Ceos, mentioned by Plato in *The Republic* (331e–332c, 335e) and in *Protagoras* (338e6–349a6) [55]. According to the Byzantine author Michael Psellos (*Tymotheos. On the Working of Demons*), Simonides of Ceos coined the phrase "the word is the image of the thing" [67], and, according to Plutarch (*On the Glory of the Athenians*), "he calls painting silent poetry and poetry painting that speaks" [65].

In terms of computational speculation on the future of communication figures and tools, the dichotomy between words and images seems to become less rigid, with a cascade of possible consequences. The advent of the computer has exponentially augmented the capacity of individual human beings of both storing and processing data. Now artificial intelligence—as an ultimate tool developed by humans (artificial, viz. human made)—is opening the doors to individuals to expand and augment human imagination. Microtext could walk this path and facilitate the merging of brachygraphy and emojis in the interest of human empathy and a more effective and direct sharing of emotions and sentiments. This could of course tremendously help to avoid the interpersonal misunderstandings that affect social media communication and also assist computational sciences in the complex processing of the polysemy of human imaging and symbolization (i.e. teaching to machines the different human understandings of reality, which involves cultural diversity). Digital semantic aggregation and visualization could paradoxically free human imagination from the mechanical philosophy constraints and expand the horizon of human creative power.

Conclusion

The amount of unstructured data has increased at an exponential rate in recent times. Handling such data has always been a challenge. Since data are also becoming noisier and noisier (but also increasingly valuable), the challenge is becoming manifold. There has been an increase in informal communication styles like SMS and chat, where the human communicator deliberately uses non-standard word forms for communication.

In this paper, we reviewed different approaches to microtext normalization. The microtext normalization technique has been classified into syntax-based, probability-based and phonetic-based approaches. Solving each of such tasks aims to solve microtext normalization from different perspectives. We also discussed available datasets, application areas

and challenges to microtext normalization. In the future, we hope to see more cognitive inspired techniques that can leverage on semantics and phonetics to take a more holistic and pragmatic approach to understand microtext the way humans do.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Informed Consent Informed consent was not required as no human or animals were involved.

Human and Animal Rights This article does not contain any studies with human or animal subjects performed by any of the authors

References

1. Agarwal S, Godbole S, Punjani D, Roy S. How much noise is too much: a study in automatic text classification. In: Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007; 2007. p. 3–12.
2. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991;6(1):37–66.
3. Baldwin T, de Marneffe M-C, Han B, Kim Y-B, Ritter A, Xu W. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In: Proceedings of the Workshop on Noisy User-generated Text; 2015. p. 126–135.
4. Bartlett S, Kondrak G, Cherry C. Automatic syllabification with structured SVMs for letter-to-phoneme conversion. Proceedings of ACL-08: HLT; 2008. pp 568–576.
5. Bayer T, Kressel U, Mogg-Schneider H, Renz I. Categorizing paper documents: a generic system for domain and language independent text categorization. Comput Vis Image Underst. 1998;70(3):299–306.
6. Beaufort R, Roekhaut S, Cougnon L-AL, Fairon C. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: ACL. Association for Computational Linguistics; 2010. p. 770–779.
7. Black A, Taylor P, Caley R, Clark R. The festival speech synthesis system. 1998.
8. Bouma G. Finite state methods for hyphenation. Nat Lang Eng. 2003;9(1):5–20.
9. Brody S, Diakopoulos N. Cooooooooooooooollllllllll-lll!!!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics; 2011. p. 562–570.
10. Cambria E, Hussain A, Havasi C, Eckl C. Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems. In: LNCS 5967; 2010. p. 148–156.
11. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intell Syst. 2017;32(6):74–80.
12. Cappelli A, Pelzer A. Dizionario di abbreviature latine ed italiane. Ulrico Hoepli. http://www.hist.msu.ru/Departments/Medieval/Cappelli. 1967.
13. Chaturvedi I, Cambria E, Welsch R, Herrera F. Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf Fus. 2018;44:65–77.
14. Choudhury M, Saraf R, Jain V, Mukherjee A, Sarkar S, Basu A. Investigation and modeling of the structure of texting language. Int J Doc Anal Recogn (IJDAR). 2007;10(3-4):157–174.
15. Chrupała G. Normalizing tweets with edit scripts and recurrent neural embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2014. p. 680–686.
16. Clark E, Araki K. Text normalization in social media: progress, problems and applications for a pre-processing system of casual english. Procedia-Soc Behav Sci. 2011;27:2–11.
17. Current RN. The Original Typewriter Enterprise 1867-1873. Wis Mag Hist; 1949. p. 391–407.
18. Daelemans W, van den Bosch A. Generalization performance of backpropagation learning on a syllabification task. In: Proceedings of the 3rd Twente Workshop on Language Technology. Enschede: Universiteit Twente; 1992. p. 27–38.
19. Daelemans W, Zavrel J, Van Der Sloot K, Van den Bosch A. Timbl: Tilburg memory-based learner. Tilburg: Tilburg University; 2004.
20. Desai N, Narvekar M. Normalization of noisy text data. Procedia Comput Sci. 2015;45:127–132. International Conference on Advanced Computing Technologies and Applications (ICACTA).
21. Doval Y, Vilares M, Vilares J. On the performance of phonetic algorithms in microtext normalization. Expert Syst Appl. 2018;113:213–222.
22. Ellen J. All about microtext-a working definition and a survey of current microtext research within artificial intelligence and natural language processing., ICAART (1) 2011; 2011. p. 329–336.
23. Fairon C, Paumier S. A translated corpus of 30,000 French SMS. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006); 2006. p. 351–354.
24. Fossati D, Di Eugenio B. I saw TREE trees in the park: how to correct real-word spelling mistakes. In: LREC. Citeseer; 2008. p. 896–901.
25. Gopalakrishna Pillai R, Thelwall M, Orasan C. Detection of stress and relaxation magnitudes for Tweets. In: Companion of the The Web Conference 2018 on The Web Conference 2018. International World Wide Web Conferences Steering Committee; 2018. p. 1677–1684.
26. Gouws S, Metzler D, Cai C, Hovy E. Contextual bearing on linguistic variation in social media. In: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics; 2011. p. 20–29.
27. Han B, Baldwin T. Lexical normalisation of short text messages: Makn sens a# twitter. In: ACL; 2011. p. 368–378.
28. Hanna S. An essential guide to singlish. Gartbooks: Singapore; 2003.
29. Hirst G, Budanitsky A. Correcting real-word spelling errors by restoring lexical cohesion. Nat Lang Eng. 2005;11(1):87–111.
30. Hocq S. Étude des sms en franćais: constitution et exploitation d'un corpus aligné SMS-langue standard. Rapport interne, Université Aix-Marseille. 2006.
31. Hoppe HR. The Third (1600) Edition of Bales's "Brachygraphy". J Engl German Philol. 1938;37(4):537–541.
32. How Y, Kan M-Y. Optimizing predictive text entry for short message service on mobile phones. In: Proceedings of HCII; 2005. vol. 5. p. 2005.
33. Jahjah V, Khoury R, Lamontagne L. Word Normalization Using Phonetic Signatures. In: Khoury R. and Drummond C, editors. Advances in artificial intelligence. Springer International Publishing; 2016. p. 180–185.
34. Jiampojamarn S, Kondrak G, Sherif T. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for

Computational Linguistics; Proceedings of the Main Conference; 2007. p. 372–379.
35. Jing H, Lopresti D, Shih C. Summarizing noisy documents. In: Proceedings of the Symposium on Document Image Understanding Technology; 2003. p. 111–119.
36. Jose G, Raj NS. Lexico-syntactic normalization model for noisy SMS text. In: 2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE). IEEE; 2014. p. 163–168.
37. Kaufmann M, Kalita J. Syntactic normalization of Twitter messages. In: International conference on natural language processing. India: Kharagpur; 2010. p. 7.
38. Khoury R. Phonetic normalization of microtext. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2015s, p. 1600–1601. 2015.
39. Kobus C, Yvon F, Damnati G. Normalizing SMS: are two metaphors better than one? In: Proceedings of the 22nd International Conference on Computational Linguistics. Vol. 1. Association for Computational Linguistics; 2008, p. 441–448.
40. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics; 2007, p. 177–180.
41. Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. p. 1137–1145. 1995.
42. Leeman-Munk S, Lester J, Cox J. Ncsu_sas_sam: deep encoding and reconstruction for normalization of noisy text. In: Proceedings of the Workshop on Noisy User-generated Text; 2015. p. 154–161.
43. Li C, Liu Y. Normalization of text messages using character- and phone-based machine translation approaches. In: Thirteenth Annual Conference of the International Speech Communication Association; 2012. p. 2330–2333.
44. Liu F, Weng F, Jiang X. A broad-coverage normalization system for social media language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers; 2012. Vol. 1. p. 1035–1044.
45. Lo SL, Cambria E, Chiong R, Cornforth D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev. 2017;48(4):499–527.
46. Lopes C, Perdigao F. Phoneme recognition on the TIMIT database. In: Speech Technologies. InTech; 2011, p. 285–302.
47. Lourentzou I, Manghnani K, Zhai C. Adapting sequence to sequence models for text normalization in social media. arXiv:1904.06100. 2019.
48. Luong M-T, Manning C. Achieving open vocabulary neural machine translation with hybrid word-character models. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016. vol. 1, p. 1054–1063. 2016.
49. Luong T, Socher R, Manning C. Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning; 2013. p. 104–113.
50. Lusetti M, Ruzsics T, Göhring A, Samardžić T, Stark E. Encoder-decoder methods for text normalization. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). Santa Fe: Association for Computational Linguistics; 2018. p. 18–28. https://www.aclweb.org/anthology/W18-3902.
51. Manning C. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: International conference on intelligent text processing and computational linguistics. Springer; 2011, pp. 171–189. 2011.
52. Miller D, Boisen S, Schwartz R, Stone R, Weischedel R. Named entity extraction from noisy input: speech and OCR. In: Proceedings of the sixth conference on Applied natural language processing. Association for Computational Linguistics; 2000, p. 316–324. 2000.
53. Mittal A, Bhatt P, Kumar P. Phonetic matching and syntactic tree similarity based QA system for SMS queries. In: 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE). IEEE; 2014, p. 1–6. 2014.
54. Mitzschke P, Lipsius J, Haffley N. Biography of the father of stenography, Marcus Tullius Tiro together with the Latin Letter De Notis. Brooklyn: Concerning the Origin of Shorthand; 1882.
55. Molyneux J. Greek Lyric, Vol. III Stesichorus, Ibycus, Simonides, and Others ed. by David A. Campbell. 1993, Vol. 37.
56. Norvig P. How to write a spelling corrector. De: http://norvig.com/spell-correct.html. 2007.
57. Peng H, Ma Y, Li Y, Cambria E. Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl-Based Syst. 2018;148:167–176.
58. Pennell DL, Liu Y. Normalization of text messages for text-to-speech. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE; 2010. p. 4842–4845.
59. Pennell D. L, Liu Y. A character-level machine translation approach for normalization of SMS abbreviations. In: IJCNLP; 2011. p. 974–982.
60. Pennell DL, Liu Y. Normalization of informal text. Comput Speech Lang. 2014;28(1):256–277.
61. Petrović S, Osborne M, Lavrenko V. The Edinburgh Twitter corpus. In: Proceedings of the NAACL HLT Workshop on Computational Linguistics in a World of Social Media; 2010. p. 25–26.
62. Pirinen TA, Hardwick S. Effects of weighted finite-state language and error models on speed and efficiency of finite-state spell-checking. In: Preproceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing FSMNLP; 2012. p. 6–14. 2012.
63. Pirinen T. A, Lindén K. State-of-the-art in weighted finite-state spell-checking. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer; 2014, p. 519–532.
64. Platt JT. The Singapore English speech continuum and its basilect'Singlish'as a'creoloid'. Anthropological Linguistics; 1975. p. 363–374. 1975.
65. Plutarch, Vol. 4. Moralia. Cambridge: Harvard University Press; 1936, p. 500.
66. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fus. 2017;37:98–125.
67. Psellu M. De operatione daemonum. A.M Hakkert; 1964. p. 2.
68. Robertson DS, et al. Phase change: the computer revolution in science and mathematics. USA: Oxford University Press; 2003.
69. Rosa KD, Ellen J. Text classification methodologies applied to micro-text in military chat. In: Proc. Eight International Conference on Machine Learning and Applications. Miami; 2009, p. 710–714.
70. Satapathy R, Guerreiro C, Chaturvedi I, Cambria E. Phonetic-based microtext normalization for Twitter sentiment analysis. In: ICDM; 2017. p. 407–413.
71. Satapathy R, Li Y, Cavallari S, Cambria E. Seq2seq deep learning models for microtext normalization. In: 2019

International Joint Conference on Neural Networks (IJCNN). IEEE; 2019.
72. Satapathy R, Singh A, Cambria E. PhonSenticNet: a cognitive approach to microtext normalization for concept-level sentiment analysis. In: CSoNet; 2019. p. 177–188. arXiv:1905.01967.
73. Schiaparelli L. Avviamento allo studio delle abbreviature latine nel medioevo. Olschki; 1926.
74. Skut W, Krenn B, Brants T, Uszkoreit H. An annotation scheme for free word order languages. In: Proceedings of the Fifth Conference on Applied Natural Language Processing. Association for Computational Linguistics, p. 88–95. 1997.
75. Taghva K, Borsack J, Condit A. Effects of OCR errors on ranking and feedback using the vector space model. Inf Process Manag. 1996;32(3):317–327.
76. Thurlow C, Brown A. Generation txt? The sociolinguistics of young people's text-messaging. Discour Anal Online. 2003;1(1):30.
77. Wang P, Ng HT. A beam-search decoder for normalization of social media text with application to machine translation. In: HLT-NAACL; 2013. p. 471–481.
78. Wilcox-O'Hearn A, Hirst G, Budanitsky A. Real-word spelling correction with trigrams: a reconsideration of the Mays, Damerau, and Mercer model. In: International conference on intelligent text processing and computational linguistics. Springer; 2008, p. 605–616. 2008.
79. Xu K, Xia Y, Lee C-H. Tweet normalization with syllables In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2015. vol. 1, p. 920–928. 2015.
80. Xue Z, Yin D, Davison BD. Normalizing microtext. Analyzing Microtext. 2011:74–79.
81. Yang Y, Eisenstein J. A log-linear model for unsupervised text normalization. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. p. 61–72. 2013.
82. Yuan S, Wu J, Wang L, Wang Q. A hybrid method for multi-class sentiment analysis of micro-blogs. In: 2016 13th International Conference on Service Systems and Service Management (ICSSSM). IEEE; 2016. p. 1–6.
83. Zhang C, Baldwin T, Ho H, Kimelfeld B, Li Y. Adaptive parser-centric text normalization. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2013. vol. 1, p. 1159–1168.