MedTxting: Learning based and Knowledge Rich SMS-style Medical Text Contraction

[1]Feifan Liu, PhD, [1]Soheil Moosavinasab, BS, [2]Thomas K. Houston, MD, [1]Hong Yu, PhD
[1]University of Wisconsin Milwaukee, Milwaukee, WI
[2]University of Massachusetts, Worcester, MA and eHealth QUERI, Bedford VA, Bedford, MA

Abstract

In mobile health (M-health), Short Message Service (SMS) has shown to improve disease related self-management and health service outcomes, leading to enhanced patient care. However, the hard limit on character size for each message limits the full value of exploring SMS communication in health care practices. To overcome this problem and improve the efficiency of clinical workflow, we developed an innovative system, MedTxting (available at http://medtxting.askhermes.org), which is a learning-based but knowledge-rich system that compresses medical texts in a SMS style. Evaluations on clinical questions and discharge summary narratives show that MedTxting can effectively compress medical texts with reasonable readability and noticeable size reduction. Findings in this work reveal potentials of MedTxting to the clinical settings, allowing for real-time and cost-effective communication, such as patient condition reporting, medication consulting, physicians connecting to share expertise to improve point of care.

Introduction

Investigation of the application of mobile computing and communication technology for improving health and health service outcomes, referred to as M-health, has been rapidly expanding[1–4]. There are now more than 5 billion mobile phone subscribers, and 90% of the world's population is covered by a cell signal[5]. In the US there are over 300 million cellular phone subscribers who send over 2.1 trillion text messages per year; almost every household in the US has at least one cellular phone and over 26% are wireless-only households[6]. In the healthcare setting, it was reported in 2011 that 72% physicians in the US use Smartphones for clinical purpose[7].

Text messaging or SMS for Short Message Service, has proved to be helpful in disease management and prevention[8], clinical and healthy behavior intervention[9], increasing clinic attendance[10], and improving health outcomes and processes of care[2]. Although these investigations demonstrate the potential of M-health, significant challenges limit the implementation of SMS mobile communication in real-world health care systems. For instance, patients and physicians still cannot get well-connected for real-time health care communications through mobile text messaging networks, although the increasing of such kinds of needs are reflected by more and more emerging web-based platforms[11–16]. Furthermore, internet might not be available in many circumstances, such as ambulance unit and urgent care practice, and many developing countries don't even have internet coverage for a large amount of areas, so SMS is a good alternative way for such communications. Therefore, it is important to identify and overcome the existing challenges in SMS based M-health, making full use of mobile communication technologies to transform how health services are delivered and change how patients and doctors interact, which can potentially lead to a great impact on global health.

Different from daily life communication vis SMS, health care-related SMS communication frequently contains complex information. One of the challenges in SMS communication for M-health is imposed by the hard 160-character limitation for each mobile short message, and even some web-based health care platforms have a character size limit[11]. Currently, mobile phones will break text messages over the limit into separate messages, and this becomes a practical concern of cost for SMS-based healthcare practices in developing countries, which not only have the majority of the world's mobile phone subscribers, but also accounts for 80% of the new ones[17,18]. Even where the cost is not a concern, arbitrary segmentations and truncations of messages are undesirable and unfriendly for users. As an example, for better communication with other colleagues regarding treatment and diagnoses, physicians need to include as much patient information (such as medication history and symptom condition) as possible, but prefer to fit those information into as few number of SMS as possible to avoid confusion and missing information. To attempt to fit character limits, users in daily life have developed messaging shortcuts (SMS lingos) to maintain the content of the message while altering spelling or phrasing to make it shorter. But it is difficult for patients/physicians to effectively and optimally apply these SMS lingos in medical texts for better health care communications. On the other hand, with full screen touch keyboards, QWERTY keyboard input, improved predictive text entry methods[19–21], and even increasingly improved speech-to-text techniques[2,22] available, full text

type speed is no longer a concern and the tediousness of text entry is decreasing. Consequently, the role of SMS lingos is greatly shifting from making typing faster towards making it easier to fit the character limit.

Therefore, the automation of medical text compression or shortening may be a valuable gain in efficiency, opening up new real-time communication scenarios between patients and physicians, among patients alike and among physicians connecting to share expertise to improve systems of care. In this paper, we present a learning-based but knowledge rich approach to automatically compress medical texts by adequately using existing SMS lingos[1] as well as predicting new lingos through the learned pattern. The fully implemented system, MedTxting, employs a statistical machine translation (SMT) learning framework enhanced by various external knowledge resources that we manually compiled with further cleaning-up and filtering. Phrase-based SMT models were trained in both word level and pronounciation level, which were finally harmonized using a heuristic method.

There have been a significant amount of work on SMS normalization in open domain[23–31]; however, to the best of our knowledge, there is no research work published on automatic compression into SMS. Findings in this work have a potential to advance network connections through SMS not only between patients and caregivers but among physicians, and reduce the costs of current M-health practices which are dependent on reimbursement from government and other health insurers[32]. Our contributions in this paper are as follows:

(1) We conducted a pilot study on automation of medical text compression using SMS lingos;

(2) We manually compiled four knowledge resources for the above task and made them available to the research community;

(3) We developed MedTxting which exploits a SMT based learning framework enhanced by existing external knowledge, and the built-in pronounciation-level model makes MedTxting robust on medical texts;

(4) We demonstrated that MedTxting can effectively compress clinical questions and discharge summary narratives in a SMS style, adequately reducing the size while keeping a reasonable readability.

Background

SMS Language Analysis

With the increasing popularity of text messaging, SMS language (also called Txt[33] or textese[34]) has been developed. Similar to chat rooms, SMS language condenses common words or sounds to allow denser messages. These linguistic adaptations aroused a lot of interest in investigating the linguistic features in SMS language[33,35] and examining its social and psychological effects within social network[36,37].

To meet the needs for many existing natural language processing applications, normalizing SMS messages has recently drawn much attention in the computational linguistic community, where the goal is to recover shortened messages into their standard English forms. SMS normalization has been handled through three well-known NLP metaphors[29]: spell checking[23–25,27,28,31], machine translation[26,30] and automatic speech recognition[29]. From the methodology point of view, they can be grouped into two categories: supervised learning method including Hidden Markov model (HMM)[24,31], machine translation (MT) [26,30], conditional random fields (CRF)[23], finite state machine[27,29], and support vector machines (SVM)[28]; and unsupervised learning method[25]. To date, there is no published work on automatic compression of texts in SMS style.

Sentence Compression

Sentence compression aims to produce a summary of a single sentence which would keep the salient content and be shorter but still grammatically correct[38]. Much of the current work typically formulates sentence expression as a word deletion problem: a shortened sentence is produced by removing any subset of the words in the input sentence[39]. Across different modeling paradigms, supervised methods include generative models[39,40] and discriminative models[39,41–43], and unsupervised methods include syntactic rule-based[44] or language model-based[45] approaches. Further studies extended the existing frameworks to allow global optimization[46], tree transduction beyond word deletion[47], and multiple sentence compression[48].

Sentence compression differs from text contraction in that the former is used to preserve syntactically salient content on the word-level and keep it grammatically sound, while the latter is to maximize the original information content with the character-level shortened expression in a "grammatically-incorrect" way. So methods and models for

[1] The main focus of this study is to explore leveraging the existing SMS lingos to facilitate medical text contraction, and the aspect of "SMS literacy" variance and its potential impact is beyond the scope of this study.

sentence compression are not applicable for SMS style compression. Similarly, data compression algorithms, such as Huffman coding[49], can't be applied in this study as the compressed version is not interpretable to users.

Methods

Knowledge Resource Compilation

We compiled four types of dictionary resources using either unsupervised rule-based approach or manual efforts.

(1) Internet SMS lingo dictionary (**Web_SMS**): We built Web_SMS by integrating a variety of internet sources related to SMS dictionary/lingo[50–53], text message shorthand[54], Twitter dictionary[55] and internet slang words[56]. Those entries sometimes are noisy, so we cleaned up all the emoticon symbols (e.g. :@ → angry), html codes (e.g. " ") and definition explanations (e.g."as in …" or "same as …"). In addition, we expanded entries with alternatives (e.g. split "abt/ab →about" into two individual entries), and removed confusing abbreviations, consisting of only numbers or combination of number and symbols (e.g. "1457→last" or "8d→manic"). Finally, we wrote tools to filter out multiple word acronyms which are more likely to be ambiguous, such as "aikrw→all I know right now". After this, we refined the original collection with 9,574 entries into one with 1829 entries.

(2) General abbreviation dictionary (**General_Abbr**): We included the list of abbreviations from the Oxford English dictionary[57] that consists of 533 entries, and added other well-recognized abbreviations, such as months, weeks, commonly used measurement units and state names of the United States. After manual review process, General_Abbr incorporates 513 entries.

(3) The UMLS abbreviations (**UMLS_Abbr**): The SPECIALIST lexicon[58] in the Unified Medical Language System (UMLS) contains 69,384 abbreviations and acronyms (2012 release). Most of these abbreviations are often used in biomedical literature, where definitions were typically provided in the same article. For this task, we filtered out this list using two criteria: the abbreviated form has only one full form association, which is assumed to be less ambiguous; there is no space in the full form, which is based on the observation that most acronyms need to be defined in advance to be interpretable. After filtering, 1904 abbreviation pairs were left in UMLS_Abbr.

(4) Clinical abbreviations (**Clinical_Abbr**): We collected medical prescription abbreviations[59] and an abbreviation list from the Clinician's Ultimate Reference[60]. After reviewing manually, Clinical_Abbr comprises of 285 entries.

Machine Learning for General SMS Compression

The goal of SMS compression is to compress the standard form English text message into a SMS-style shortened version. Similar to previous work[26,30] on normalization of SMS messages, we formulated this task as a standard statistical machine translation (SMT) problem where different translation patterns can be learned from the training data and applied on unseen data.

Machine translation model is based on the noisy channel model. Given an input of a standard English sentence E, the goal is to compress it into a SMS style word sequence S. Using Bayes rule, it is equivalent to finding the sequence S that maximizes the following:

$$\hat{S} = \arg\max_S p(S \mid E) = \arg\max_S p(E \mid S) \times p(S) \quad (1)$$

This allows for a language model for $p(S)$ and a translation model of $p(S|E)$.

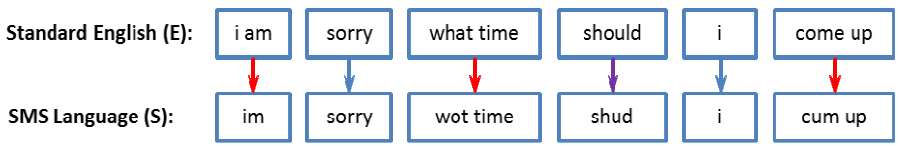

Figure 1. Illustration of phrased-based SMT for general SMS compression

We employed the phrase-based word-level SMT using the state-of-the-art open source Moses toolkit[61]. During the decoding stage, an input English sentence E is segmented into a sequence of consecutive words (so-called "phrases"), and each of them is translated into a candidate SMS phrase. The output would be optimized by both the phrase-based translation model of $p(E|S)$ trained on parallel E-S sentence pairs and the language model of $p(S)$ trained on sentences written in SMS lingos. Figure 1 illustrates an example of one E-S sentence pair in the phrase-

based machine translation, where different colors indicate different types of translations. Note that unlike other translation tasks, we don't need to consider reordering for this task.

MedTxting: Medical Texts Compression in a SMS Style

Compared with parallel SMS data from the general domain, the parallel medical text and medical SMS pairs are more expensive to obtain. Thus we cannot train a SMS compression model directly in the medical domain as we can for the general domain. To solve this problem, we further developed MedTxting to extend the general domain model by incorporating various external knowledge resources. The assumptions are (1) there is a core set of abbreviations that can be used across varying SMS linguistic regions[30] and (2) SMS lingo patterns can be learned through pronounciation-level SMT model, which is more generalizable than word-level model[29]. Figure 2 shows the system diagram of MedTxting. There are four main modules in MedTxting, which we will describe in more detail.

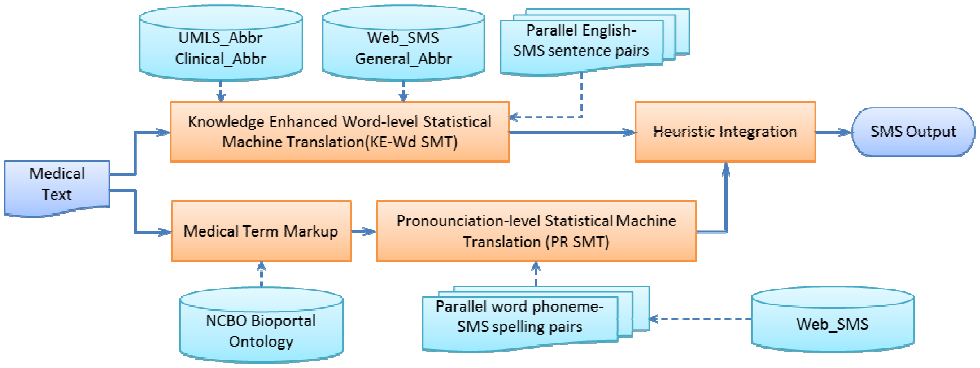

Figure 2. System Diagram of MedTxting.

(I) Knowledge Enhanced Word-level (KE-Wd) SMT Module

This module is built based on the general model (as described in last section) trained on English-SMS sentence pairs as shown in Figure 2. The hypothesis is that most knowledge learned from general SMS parallel data can be portable to medical domain, e.g. translation knowledge in general is also applicable in medical domain. One weakness would be the role of language model in the optimization process (Eq. (1)) becomes less effective due to the context variation across different domains. In this case, we used external knowledge resources to provide more guidance over the translation model to make up this weakness.

For the four dictionaries plugged in the statistical machine translation system as shown in Figure 2, we assigned a larger weight of "1.0" to UMLS_Abbr and Clinical_Abbr than the weight of "0.8" to Web_SMS and General_Abbr. The reason for this is that Web_SMS and General_Abbr share some knowledge with the parallel corpus and we would like for them to be more coordinated.

(II) Medical Term Markup Module

As a preprocessing module to the pronounciation-level SMT described later, the medical term markup module tries to protect medical domain-specific terms from being contracted unless they are found in UMLS_Abbr or Clinical_Abbr. The purpose for this is to minimize the compression on these terms with clinically significant meaning, such as medication, disease and symptoms.

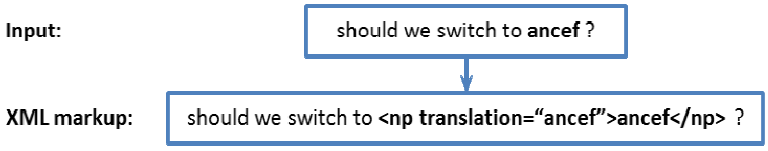

Figure 3. An example of medical term markup output.

To do that, we use the open biomedical annotator[62] web service API developed by the National Center for Biomedical Ontology (NCBO). The ontologies we used for this module are RxNorm, SNOMED Clinical Terms (SNOMED CT), MedDRA, International Classification of Diseases version 9 (ICD-9), Human disease (DOID) and Chemical entities of biological interest (CHEBI). We also restricted the annotation to the following semantic types defined in UMLS: Disease or Syndrome (T047), Finding (T033), Sign or Symptom (T184), Inorganic Chemical (T197), Neoplastic Process (T191), Organic Chemical (T109), Pharmacologic Substance (T121), Steroid (T110) and

Substance (T167). Terms annotated are marked-up using xml format, as shown in Figure 3, to notify downstream modules that the translation has been specified.

(III) Pronounciation-level(PR) SMT Module

We observed that many SMS style compressions are achieved by pronounciation similarity. For example, the phoneme sequence of "ae t" (pronounciation of "at") is often replaced by "@" (e.g. "battery→b@re"); "ey t" (pronounciation of "eight") is often replaced by "8" (e.g. "straight → str8"). Phonemes representing pronounciation are a close set and much more generalizable than words, thus SMS compression patterns based on phonemes learned from the general domain is more portable on medical texts.

Motivated by that, we trained another phrase based SMT model on pronounciation-level instead of aforementioned word-level. Specifically, given an input English word represented by a sequence of its phonemes *P*, the goal is to compress it into a SMS style letter sequence *L* (predicted lingo). Training pairs for this model would be parallel phoneme-letter sequence pairs, e.g. for the compression word pair "atmosphere←→@mosfer", the training pair would be "ae t m ax s t f iy r"←→ "@ m o s f e r". Similarly, Eq. (1) would be transformed as follows:

$$\hat{L} = \arg\max_L p(L \mid P) = \arg\max_L p(P \mid L) \times p(L) \tag{2}$$

(IV) Heuristics-based Integration Module

Heuristics-based Integration

Let *O* be the original word sequence
W be the word sequence from word model
P be the word sequence from pronounciation model
M be the set of marked-up medical terms
for *i* from 0 to length(*W*)-1 do
if (length(w_i) > 5) AND ($w_i \in O$) AND ! ($w_i \in M$)
if (length(p_i) < length(w_i))
$w_i = p_i$;
end if
end if
end for

Figure 4. Heuristic rule in the integration module.

Theoretically the word-level SMT model will be conservative due to more contextual constraints and it needs more annotated data for better coverage, while the pronounciation level model will be aggressive due to its contraction ability for almost each word. How to obtain an optimized balance between contraction ability and readability is a challenging task. In the current version of MedTxting, we use a simple heuristics-based method to attempt to integrate two models together, where the pronounciation model is applied only if the output word from word-level model has a larger character size of 5, has not been contracted yet, and is not marked-up as a medical term(See Figure 4 for details).

Results

Data and Experiment Setup

We use the corpus from Raghunathan et al.[30] which was created based on the subset of NUS corpus[63], HKU corpus[64], TMT corpus[31] and corpus from Aw et al[26]. The data consists of 9272 parallel standard English-SMS pairs. To develop a general SMS contraction model, we split the data into a set of 6490 pairs for training, a set of 1854 pairs for parameter tuning(development set), and another set of 928 pairs for testing. For MedTxting, we used all the available SMS pairs to train the word-level SMT model, incorporating 2 medical knowledge resources and 2 general dictionaries we compiled for this task. Among the four knowledge resources, we chose Web_SMS to generate the training data for the pronounciation model, where for each full form word we obtained its phonemes using the NIST standard text-to-phone tool[65], and paired with the letter sequence of its SMS counterpart for model training. To test MedTxting on medical texts, we randomly chose 10 clinical questions[66] and 10 de-identified discharge summary narratives[67] with a character length range from 200 to 300, and the average number of words (including punctuations and symbols) per medical text is 43.25.

We used standard state-of-the-art open source tools to train a phrase-based either word-level or pronunciation-level machine translation model. SRI language modeling toolkit was used for training a language model and GIZA++ for computing alignments between counterparts of standard English and SMS messages. Finally Moses[61] was used to

train SMT models capable of decoding from standard English to SMS-style contractions, where we used minimum error rate training (MERT)[68] for parameter tuning on the development dataset. The general SMS contraction system was evaluated using standard BLEU metric[69]. Due to a lack of gold standards for medical SMS contraction, we asked three physicians to recover the original text given the compressed text from MedTxting. We defined a metric, called Correctly Recover Rate (CRR), to calculate the percentage of unigrams and bigrams that can be correctly recovered from the compression as follows.

$$CRR\_unigram = \frac{\#of\ \ unigrams\ \ correctly\ \ recovered}{\#of\ \ unigrams\ \ changed\ \ by\ \ contraction} \qquad (3)$$
$$CRR\_bigram = \frac{\#of\ \ bigrams\ \ correctly\ \ recovered}{\#of\ \ bigrams\ \ changed\ \ by\ \ contraction}$$

In addition, we evaluated the system's contraction ability by the contraction ratio calculated by the character number reduced by the system divided by the total character number in the original messages.

Performance of SMT-based Compression on General SMS Corpus

We first examined four general SMS compression system settings on the testing data set. (1) **Dict:** a dictionary based system to look up each word in Web_SMS and General_Abbr and replace them with their counterparts in the dictionary. Note that for each full form text token there might be multiple SMS candidates, and for this experiment we simply used the first one in the alphabetic order. (2) **SMT**: word-level SMT system trained on training data and tuned on development data. (3) **SMT+Dict (hard)**: incorporates Web_Abbr and General_Abbr in an exclusive way, which means the SMT system will fully respect the dictionary match. (4) **SMT+Dict (soft)**: incorporates two dictionaries in an inclusive way, which means the dictionary match is treated as a candidate to compete with others in the SMT decoding process.

Table 1. Automatic compression performance comparison between different systems on general SMS data.

	BLEU	*BLEU unigram*	*BLEU bigram*	*Contraction Ratio*
Dict	0.3667	0.653	0.443	13.38%
SMT	0.707	0.860	0.75	7.07%
SMT+Dict(hard)	0.401	0.677	0.473	14.08%
SMT+Dict(soft)	0.61	0.802	0.662	10.39%

As shown in Table 1, SMT achieved the best BLUE score of 0.707, but the worst contraction ratio of 7.07%. By incorporating dictionary resources, the contraction ratio was increased to 10.39% (inclusive) and 14.08% (exclusive) respectively, at the cost of decreasing BLEU score. Dictionary only (Dict) obtained the lowest BLEU but a nice contraction ratio of 13.38%. We observe that the contraction ratio on the training data and testing data is 8.53% and 8.34%, explaining that on this data collection people don't seem to use as many contractions as possible, partially because a lot of them are short quick-updating messages. So Dict will aggressively search and replace possible contractions, resulting in larger contraction ratio but lower BLEU score.

Interestingly, compared with Dict, SMT+Dict (hard) improved both the BLEU score and contraction ratio, showing that in addition to un-comparable contextual information the SMT model also complements with dictionary resources on the translation knowledge. Based on the BLEU overall score and unigram score, SMT+Dict (soft) would be a better solution for real-world applications with the BLEU unigram score of 0.802 and contraction ratio of 10.39%. BLEU bigram scores show a similar pattern with unigrams and overall BLEU scores across different systems.

Evaluation of MedTxting on Medical Narratives

In this section we will evaluate MedTxting's performance on medical texts. Based on findings in the above experiments, we set up the knowledge-enhanced word-level SMT module using inclusive ways to incorporate four dictionary resources into SMT. We assigned "Clinical_Abbr" and "UMLS_Abbr" a weight of 1, "Web_SMS" and "General_Abbr" a weight of 0.8. Three settings were evaluated: **Dict_Med** for a dictionary lookup system using all four dictionary resources; **MedTxting** for the setting containing all the modules in Figure 2; **MedTxting w/o PR** for the setting excluding the pronounciation model. Results are in Table 2.

Table 2. MedTxting evaluation on clinical questions and discharge summary narratives

	CRR_unigram	*CRR_bigram*	*Contraction Ratio*
Dict_Med	67.85%	63.45%	11.47%
MedTxting	57.52%	51.11%	18.73%
MedTxting w/o PR	72.22%	65.39%	10.14%

We can see that MedTxting w/o PR achieved 72.22% of CRR_unigram and 65.39% CRR_bigram, outperforming both MedTxting and Dict_Med. On the other hand, with the help of PR SMT model, MedTxting obtained the best contraction ability at the contraction ratio of 18.73%, compared with the Dict_Med of 11.47% and MedTxting w/o PR of 10.14%. For the three metrics, Dict_Med stands in the middle. To gain a better understanding of how the contraction works, Table 3 demonstrated the compression outputs from three systems on a sample clinical question.

Table 3. Demonstration of automatic compression outputs on a sample medical text (contractions are highlighted in red)

System/Output	A Sample Clinical Question
Original Text	how to sort this out ? complaining of right arm pain and numbness - is it the zoster scarring or is he developing carpal tunnel syndrome ? also he is diabetic and lipids are elevated so that could also be a factor .
Dict_Med	how **2** sort **dis** out ? **c/o r8** arm pain **&** numbness - is it **da** zoster scarring or is he developing carpal tunnel syndrome ? also he is diabetic & lipids **r** elevated so **dat cld** also **b** a factor .
MedTxting	how **2** sort **dis** out ? **c/o r8** arm pain **&** numbness - is it **da** zoster **scrin** or is he **divelpn** carpal tunnel syndrome ? also he is diabetic & lipids **r luv8d** so that could also **b** a **factr** .
MedTxting w/o PR	how **2** sort **dis** out ? **c/o r8** arm pain **&** numbness - is it **da** zoster scarring or is he developing carpal tunnel syndrome ? also he is diabetic **&** lipids **r** elevated so that could also **b** a factor .

As we can see, the three systems share some common knowledge on contractions. However, sometimes well-accepted contractions in the general SMS domain might not be recognized by physicians, such as "this→dis" or "the→da" which is used by all three systems. We notice that with pronounciation model, MedTxting can predict adequate contractions that SMT itself or dictionary cannot, e.g. "developing → divelpn" and "factor → factr". Through some error analysis, we observed that for the pronounciation model, the first letter or first phoneme should be kept for a better contraction. For example, in the question above "elevated" is contracted to "luv8d", which is actually a very good contraction if the first letter was kept as "eluv8d". In addition, we found that the threshold of 5 in character length in the integration heuristics might not be a good choice as many adequate contractions were blocked from the pronounciation model, such as "that →th @" and "could → c%d".

Discussion

This study shows the potential of our approach for integrating automatic text contraction applications into M-health platforms so that physicians and patients are better connected through real-time healthcare communications, wherever there is a cell signal covered. The contraction function can at least be able compress the message size 10.14% less (based on this pilot study); representing a great cost reduction for SMS based healthcare interventions which have been dependent on insurance reimbursement.

There is much room for further improvement on each module of MedTxting and the integration methodology. With the external knowledge resources (SMS lingos), we have shown in this study that MedTxting has achieved a relatively adequate performance in automatically contracting medical texts. Although our work does not depend upon any annotated data from the medical domain, we speculate that the MedTxting performance can be further improved when in-domain annotated clinical data and updated knowledge resources are available. However, it is expensive to create annotated data, an alternative is to explore un-annotated monolingual corpus to improve the system's performance, as proven in Liu et al[70] on the machine translation task.

The markup model is not effective enough to protect some clinically significant words from contraction while at the same time, falsely marks up non-clinical generic terms, e.g. "pain" and "medicine." On the other hand, the model fails to shorten many medical terms as physicians typically do. For example, a physician would shorten the example

in Table 3 as "how to sort? R arm pain/loss ses'n – VSV or CTS? PT is DM's hyperlipid, contributing?" in which "DM" represents "diabetic" and "zoster" and "carpal tunnel syndrome" are abbreviated as "VZV" (for varicella-zoster-virus) and "CTS," respectively.

One innovation in our work is the pronunciation model, which can infer associations between phonemes patterns and contraction patterns, specifically related to digits or special symbols. However, the model is character based, lacking contextual constraints, and resulting in a poor readability. We empirically assigned some parameters and heuristics in our model. We speculate that a learning based re-ranking approach leveraging probability of each model would be an optimized way for integration.

Conclusions and Future Work

We conducted a pilot study on automatic SMS contraction, presented and evaluated a learning based and knowledge rich method on both the general SMS domain and the medical domain. Our experimental results show that SMT based model and knowledge resources can effectively interact with each other without the necessity of a parallel in-domain annotated data. The developed MedTxting system demonstrated promising adequacy in clinical texts with regard to correct recover rate and contraction ratio. The survey with the evaluation also shows that more than half of contracted messages (55%) from MedTxting w/o PR were checked by physicians showing the willingness to send this type of SMS to physician colleagues for seeking clinical advice and other activities related to patient treatments, assuming that the SMS message exchange is allowed in their hospitals and there is absolutely no privacy and security concerns.

For future work, we plan to incorporate a letter-transformation model[23] to improve MedTxting's robustness, and explore in-domain clinical data to boost performance of statistical translation for automatic SMS contraction. The current evaluation is based on a small sample of medical texts, and we will conduct more extensive evaluations of MedTxting on a larger data set with different types of clinical narratives, as well as the how effective it supports real world communications in the clinical setting. Finally, we will investigate the contribution and property of each module to optimize the systematic integration.

Acknowledgments

The authors like to thank Steven Belknap, Robert Scott and Kourosh Rawaz for their participation in the evaluation of this study. This work is supported by the grant 1R01GM095476.

References

1. Vodopivec-Jamsek V, de Jongh T, Gurol-Urganci I, et al. Mobile phone messaging for preventive health care. In: *The Cochrane Library*. John Wiley & Sons, Ltd; 2008. Available at: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007457/abstract. Accessed March 12, 2012.
2. Krishna S, Boren SA, Balas EA. Healthcare via cell phones: a systematic review. *Telemed J E Health*. 2009;15(3):231–240.
3. Free C, Phillips G, Felix L, et al. The effectiveness of M-health technologies for improving health and health services: a systematic review protocol. *BMC Res Notes*. 2010;3:250.
4. Bäck I, Mäkelä K. Mobile Phone Messaging in Health Care–Where are we Now. *J Inform Tech Soft Engg*. 2012;2(106):2.
5. Anon. How Text Messages Could Be the Future of Healthcare - Popular Mechanics. Available at: http://www.popularmechanics.com/science/health/med-tech/how-text-messages-could-change-global-healthcare. Accessed March 12, 2012.
6. Stenner SP, Johnson KB, Denny JC. PASTE: patient-centered SMS text tagging in a medication management system. *J Am Med Inform Assoc*. 2011. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21984605. Accessed March 10, 2012.
7. Anon. 72 percent of US physicians use smartphones. *Mobihealthnews.com*. 2010. Available at: http://mobihealthnews.com/7505/72-percent-of-us-physicians-use-smartphones/. Accessed June 8, 2011.
8. Cole-Lewis H, Kershaw T. Text Messaging as a Tool for Behavior Change in Disease Prevention and Management. *Epidemiol Rev*. 2010;32(1):56–69.
9. Wei J, Hollin I, Kachnowski S. A review of the use of mobile phone text messaging in clinical and healthy behaviour interventions. *J Telemed Telecare*. 2011;17(1):41–48.
10. Guy R, Hocking J, Wand H, et al. How Effective Are Short Message Service Reminders at Increasing Clinic Attendance? A Meta-Analysis and Systematic Review. *Health Services Research*. 2012;47(2):614–632.
11. Anon. HealthTap. Available at: https://www.healthtap.com/. Accessed March 13, 2012.
12. Anon. Parkhurst Exchange. Available at: http://www.parkhurstexchange.com/. Accessed March 13, 2012.
13. Anon. MedHelp. Available at: http://www.medhelp.org/. Accessed March 13, 2012.

14. Anon. Online Support Groups and Forums at DailyStrength. Available at: http://www.dailystrength.org/. Accessed March 13, 2012.
15. Anon. Ask a Patient: Medicine Ratings and Health Care Opinions. Available at: http://www.askapatient.com/. Accessed March 13, 2012.
16. Anon. Clinicians Turn to Twitter to Connect and Share Expertise. Available at: http://www.asha.org/Publications/leader/2011/111122/Clinicians-Turn-to-Twitter-to-Connect-and-Share-Expertise/. Accessed March 13, 2012.
17. Déglise C, Suggs LS, Odermatt P. Short Message Service (SMS) Applications for Disease Prevention in Developing Countries. *Journal of Medical Internet Research*. 2012;14(1):e3.
18. Fontelo P, Liu F, Muin M, Tolentino H, Ackerman M. Txt2MEDLINE: Text-Messaging Access to MEDLINE/PubMed. *AMIA Annu Symp Proc*. 2006;2006:259–263.
19. Shieber SM, Nelken R. Abbreviated text input using language modeling. *Natural Language Engineering*. 2006:1.
20. Pennell DL. An improved system for text entry on cell phones. 2008.
21. Nesbat SB. A system for fast, full-text entry for small electronic devices. In: *Proceedings of the 5th international conference on Multimodal interfaces*.; 2003:4–11.
22. Anon. Pushing the Limits of Google's Speech Recognition - NYTimes.com. Available at: http://gadgetwise.blogs.nytimes.com/2009/06/29/pushing-the-limits-of-googles-speech-recognition/. Accessed March 12, 2012.
23. Liu F, Weng F, Wang B, Liu Y. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. *Proc. of ACL-HLT*. 2011.
24. Pennell D, Yang Liu. Toward text message normalization: Modeling abbreviation generation. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2011:5364–5367.
25. Cook P, Stevenson S. An unsupervised model for text message normalization. In: *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*.; 2009:71–78.
26. Aw AT, Zhang M, Xiao J, Su J. A phrase-based statistical model for SMS text normalization. In: *Proceedings of the COLING/ACL on Main conference poster sessions*.; 2006:33–40.
27. Beaufort R, Roekhaut S, Cougnon LA, Fairon C. 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*.; 2010:770–779.
28. Han B, Baldwin T. Lexical normalisation of short text messages: Makn sens a# twitter. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*.; 2011.
29. Kobus C, Yvon F, Damnati G. Normalizing SMS: are two metaphors better than one? In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*.; 2008:441–448.
30. Raghunathan K, Krawczyk S. *Investigating sms text normalization using statistical machine translation*. Stanford University, Stanford, CA; 2009.
31. Choudhury M, Saraf R, Jain V, et al. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*. 2007;10(3):157–174.
32. Adler R, Foundation CH. *Health care unplugged: The evolving role of wireless technology*. California Healthcare Foundation; 2007.
33. Tagg C. A Corpus Linguistics Study of SMS Text Messaging. 2009.
34. Drouin M. College students' use of text message abbreviations and relations with literacy. *Journal of Computer Assisted Learning*. 2011;27. Available at: http://opus.ipfw.edu/psych_facpubs/51.
35. Grinter RE, Eldridge MA. y do tngrs luv 2 txt msg? In: *ECSCW 2001*.; 2002:219–238.
36. Reid D, Reid F. Insights into the social and psychological effects of SMS text messaging. *Text*. 2004;2005(February):1–11.
37. Alison Bryant J, Sanders-Jackson A, Smallwood AMK. IMing, text messaging, and adolescent social networks. *Journal of Computer-Mediated Communication*. 2006;11(2):577–592.
38. Jing H. Sentence reduction for automatic text summarization. In: *Proceedings of the sixth conference on Applied natural language processing*.; 2000:310–315.
39. Knight K, Marcu D. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*. 2002;139(1):91–107.
40. Galley M, McKeown K. Lexicalized Markov grammars for sentence compression. *the Proceedings of NAACL/HLT*. 2007:180–187.
41. McDonald R. Discriminative sentence compression with soft syntactic evidence. In: *Proceedings of EACL*.; 2006:297–304.
42. Le Nguyen M, Shimazu A, Horiguchi S, Ho BT, Fukushi M. Probabilistic sentence reduction using support vector machines. In: *Proceedings of the 20th international conference on Computational Linguistics*.; 2004:743.

43. Riezler S, King TH, Crouch R, Zaenen A. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*.; 2003:118–125.
44. Lin J, Wilbur WJ. Syntactic sentence compression in the biomedical domain: facilitating access to related articles. *Information Retrieval (in press)*. 2007.
45. Hori C, Furui S. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E SERIES D*. 2004;87(1):15–25.
46. Clarke J, Lapata M. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*. 2008;31(1):399–429.
47. Cohn T, Lapata M. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*. 2009;34:637–674.
48. Filippova K. Multi-sentence compression: Finding shortest paths in word graphs. In: *Proceedings of the 23rd International Conference on Computational Linguistics*.; 2010:322–330.
49. Huffman DA. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*. 1952;40(9):1098–1101.
50. Anon. Popular Texting TxT Lingo list. Available at: http://www.lingo2word.com/lists/txtmsg_listA.html. Accessed March 14, 2012.
51. Anon. LG Mobile Phones: DTXTR Glossary. Available at: http://www.dtxtrapp.com/glossary.htm. Accessed March 14, 2012.
52. Anon. SMS lingo dictionary | Text SMS Messages. Available at: http://www.newmobilemedia.com/sms-lingo-dictionary.htm. Accessed March 14, 2012.
53. Anon. SMS Dictionary - Abbreviations. Available at: http://smsdictionary.co.uk/abbreviations. Accessed March 14, 2012.
54. Anon. The Largest List of Text Message Shorthand (IM, SMS) and Internet Acronyms. Available at: http://www.netlingo.com/acronyms.php. Accessed March 14, 2012.
55. Anon. Twittonary | A Twitter Dictionary. Available at: http://twittonary.com/. Accessed March 14, 2012.
56. Anon. Internet Slang words - InternetSlang.com. Available at: http://www.internetslang.com/list.asp?i=all. Accessed March 14, 2012.
57. Anon. The Oxford English Dictionary: List of Abbreviations. Available at: http://www.indiana.edu/~letrs/help-services/QuickGuides/oed-abbr.html. Accessed March 14, 2012.
58. McCray AT. The nature of lexical knowledge. *Methods Inf Med*. 1998;37(4-5):353–60.
59. Anon. List of abbreviations used in medical prescriptions - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/List_of_abbreviations_used_in_medical_prescriptions. Accessed March 14, 2012.
60. Anon. Common medical abbreviations. Available at: http://www.globalrph.com/abbrev.htm#A. Accessed March 14, 2012.
61. Koehn P, Hoang H, Birch A, 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*.; 2007:177–180.
62. Jonquet C, Shah NH, Musen MA. The Open Biomedical Annotator. *Summit on Translat Bioinforma*. 2009;2009:56–60.
63. How Y, Kan MY. Optimizing predictive text entry for short message service on mobile phones. In: *Proceedings of HCII*.Vol 5.; 2005.
64. Adams Bodomo. *Research Project on Linguistic Features of Mobile Communication.* Department of Linguistics, HKU; 2002.
65. Fisher WM. A Statistical Text-To-Phone Function Using Ngrams And Rules. *ICASSP*. 1999;2:649–652.
66. NLM. *The ClinicalQuestion Collection, http://clinques.nlm.nih.gov/JitSearch.html*. 2009.
67. Anon. University of Pittsburgh NLP Repository. Available at: http://www.dbmi.pitt.edu/blulab/nlprepository.html. Accessed March 15, 2012.
68. Och FJ. Minimum error rate training in statistical machine translation. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*.; 2003:160–167.
69. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002:311–318. Available at: http://dx.doi.org/10.3115/1073083.1073135. Accessed March 15, 2012.
70. Liu Z, Wang H, Wu H, Li S. Improving Statistical Machine Translation with Monolingual Collocation. 2010;(July):825–833.