an Article from

Journal of Advanced Computational Intelligence and Intelligent Informatics

4F Toranomon Sangyo Bldg., 2-29, Toranomon 1-chome, Minatoku, Tokyo 105-0001, Japan
Tel. +813-3508-0051, Fax: +813-3592-0648, E-mail: jaciii@fujipress.jp
homepage URL: http://www.fujipress.jp/JACIII/

Paper:

Turing Test-Based Evaluation of an Experimental System for Generation of Casual English Sentences from Regular English Input

Eleanor Clark and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
E-mail: {eleanor, araki}@media.eng.hokudai.ac.jp
[Received August 1, 2012; accepted September 24, 2012]

This paper proposes an experimental system for generating slang-style casual English sentences from regular English input using a phonetic database approach, primarily as an AI task, with real-life applications such as social media marketing. An original database consisting of multiple candidates of casual English phonemes was constructed, and linguistic analysis of Twitter data used to establish the optimum frequency of slang tokens per sentence. The human-likeness and legibility of output sentences of the experimental system were evaluated using an experiment based on the classical definition of the Turing test, in which fifty human evaluators attempted to distinguish sentences produced by the system from genuine human-authored sentences. The experiment results demonstrated that the gap in human-likeness scores between the "human" and "machine" sentences was small, and that some "machine" sentences actually outperformed several of the "human sentences." The "machine" sentences' average score of 3.1 on a 5-point scale, where 3 indicated complete uncertainty of whether the sentences were human-authored or machine-authored, can be considered a pass of the Turing test in the established definition. In this paper, we describe the potential approaches to the task, the construction of the phonetic database and the proposed system, and discuss the evaluation results.

Keywords: natural language processing, slang generation, artificial intelligence, Turing test, Twitter

1. Introduction

The proliferation of highly irregular casual written English in electronic communications including emails, chat applications, SMS (Short Message Service, mobile phone text messages), and microblogs such as Twitter[1] has created a large volume of publicly available data, but the irregularity and creativity of the language poses a problem for Natural Language Processing (NLP) applications such as Machine Translation, Information Retrieval, ontology creation, and summarization [1, 2]. In addition to the obvious problem that such text may be difficult to understand for general readers, another issue is that it is not always straightforward for such users to actually devise creative graphemes themselves and place them appropriately in their written text. Creating convincing colloquial language can be seen as a highly difficult task, as it can be considered to fall into the sphere of the Turing test [3].

We aimed to design a system that could produce credibly natural slang-like text from normal language; i.e., convert regular English input into casual English output automatically. In this method, we utilized a phoneme-by-phoneme approach, which attempts to mimic SMS or Twitter-style phonetic spellings by selecting replacement candidates at the phoneme level. As this method can produce highly creative phonetic slang, it is necessary to strike an appropriate balance between "interesting" and "difficult to understand," in which lies the difficulty of the task. Thus, this topic can be located within the scope of Artificial Intelligence (AI), since it aims to replicate human creativity. It should be clarified that this approach does not attempt to generate content itself as a chatbot or other application does, but to convert regular English to casual English in a creative way.

Regarding application of such a system, we propose that automatic generation of slang-type English from regular input text would be useful in areas such as social media marketing, targeting teenage consumers, enabling older users to communicate more smoothly with younger users in applications such as Twitter or SMS, or making chatbots seem more humanlike.

2. CEGS: A Casual English Generation System

2.1. Approaches to Casual English Generation

In our previous research on the task of casual English normalization, we used a token-to-token database (broadly speaking, word-to-word, although phrase-to-phrase of any number was also possible) for accuracy [1]. However, it is debatable whether a token-to-token database would be most appropriate for a genera-

1. http://twitter.com

tion system. The goal here is *creativity* rather than *accuracy*; if all words are converted in the same way in each sentence, the humanlike creativity aspect may be weakened. In humans, five different people may write the same word in five different ways (e.g., "this" could be written as *dis, diss, diz, thiz, viss*); thus, it may not be interesting in AI terms to use a dictionary-lookup style which states that, for example, "this" must always be converted to *diss*. However, a database which has multiple candidates for common words and a random selection algorithm would negate this problem somewhat.

Another problem with any kind of approach which relies solely on token-to-token database lookup is that out-of-vocabulary (OOV) items could not be converted. A learning method for new words using web-based searches would need to be added in order to keep the database relevant in the face of quickly-evolving language and new slang coinage.

An alternative database approach would be a phoneme-by-phoneme approach, which would mimic SMS (short message service) or Twitter-type phonetic spellings by selecting replacement candidates at the phoneme level. This would be useful for two reasons: by using phonemes, it removes the problem of OOV completely; and as a method which attempts to mimic the process of casual English token creation from scratch rather than use a pre-created database, it may be regarded as a more interesting approach from an AI standpoint. However, a disadvantage of a phoneme-to-phoneme approach is that it will lose the variety of non-phonetic casual English: acronyms, slang etc., as these are not usually based on pronunciation. Furthermore, it would face similar problems to text-to-speech applications when heteronyms appear: e.g., should the word "read" be converted to "reed" or "redd"? Depending on context, both are possible:

Did you read that book? >*Did u reed dat buk?*
Yes, I read that book. >*Yeah i redd dat buk.*

In order to tackle the latter problem, we hypothesized that the utilization of speech synthesis technology would be an effective strategy.

Statistical approaches have been utilized in normalizing slang to regular English [4, 5], so it seems logical to assume that a statistical method would also be useful in a reverse system of regular English to slang. However, a major disadvantage of this approach is that large-scale parallel corpora of casual English sentences with manually normalized regular English counterparts need to be built as the base for the SMT-like (Statistical Machine Translation) system, which is a non-trivial task. For example, Aw et al. manually normalized a substantial data set of 5,000 raw SMS messages, yet still found that OOV posed a considerable problem when words appeared which did not occur as casual English with their manually normalized equivalents in the parallel corpora [4]. Thus, despite requiring a labor-intensive creation of parallel corpora, such an approach would remain limited in terms of producing completely new coinages.

Therefore, we proposed the application of a method using a phoneme-to-phoneme approach to the task of generation of casual English sentences from regular English input. In order to tackle the pronunciation-based problem described above, we utilized the public domain resource of the CMU Pronouncing Dictionary [6], which was developed by Carnegie Mellon University primarily for use in speech synthesis, and features a mapping of over 130,000 English words to phonetic representation using a set of 39 phonemes. In the proposed method, selected tokens are split into phonemes using the CMU Pronouncing Dictionary, and these phonemes are then converted into the multiple alternative phonemes in a specifically designed database. The experimental system we developed based on this method is CEGS, a Casual English Generation System. This system is original in that, to the authors' knowledge, no other research exists on the same topic.

2.2. Casual Token Frequency per Sentence

One important point in casual English sentence design is that, usually, not all tokens (words) in a given sentence are irregular. As a very broad generalization, it appears that only a small proportion of tokens per sentence tend to be casual English items (this may, however, often be enough to render the sentence incomprehensible to a non-native speaker or to a Machine Translation application [1]).

The ultimate goal for CEGS is to be a natural recreation of human "slangification" of regular English input sentences. There are several questions that arise regarding this aim. How should the system select which words to convert in a sentence? Would an extremely simple rule such as "convert 25% of all tokens at random incidence" or even "convert every fourth token" create an impression of humanlike creativity? Or are there particular parts of speech (POS) which are more likely to be converted, e.g., are nouns more commonly written in slang than verbs? Before making the first steps in designing a method of casual English generation, these rules must be clarified. We proposed to devise these rules based on empirical data; as such, we conducted a preliminary linguistic analysis experiment on 4,716 words (320 tweets) from Choudhury's Twitter corpus [7]. There were two factors for analysis: first, we attempted to extrapolate the average occurrence per sentence (which we defined as "AOpS") of casual English items. Second, we attempted to establish which, if any, parts of speech (POS) were particularly likely to be written in casual English. The aim of determining these two points was for the proposed system to be capable of mimicking human casual English creation as naturally as possible, by recreating the most commonly seen frequency and distribution trends.

For the purpose of this experiment, casual English items were defined as a) tokens (words) which were flagged by the open source spellchecker Hunspell, *and* b) not named entities or obvious unintentional typing/spelling errors (e.g., a): *Bieber,* b): *appology*). As

some clearly intentional spelling errors are typical casual English items used for brevity or style reasons (e.g., *whateva, nuffing*), any spelling error which was deemed to have a likelihood of being intentional was included in the casual English category.

The tweets were taken from the most recent section ("Fall 2009") of Choudhury's Twitter corpus. Although selection of tweets for data was again essentially random, tweets written in languages other than English and tweets written in entirely standard English (while mostly English, the corpus also contains tweets in German and Spanish, and others) were excluded from the experiment data. The latter was due to the fact that the aim of this experiment was to analyze the construction of sentences which feature casual English items, not to analyze the incidence of casual English in Twitter sentences as a whole.

POS were recorded by manual annotation, as a conventional POS tagger cannot function effectively on such noisy text. The POS categories were noun, verb, pronoun, adverb, preposition, conjunction, interjection, and contraction. Although the first eight represent traditional English POS categories, contraction was added for the purpose of this experiment due to its frequent occurrence. Contraction refers to contractions of any pair or greater number of tokens which have been written as one token. For example, *imma* (*I'm going to*). Emoticons, though occurring with relative frequency, were not included as a POS category and thus were not counted as casual English in this experiment. Other tokens manually stripped from the data were URLs and usernames. An example sentence with manual POS tagging is shown below.

Welcome 2 Valencia, Spain! once the weather settles dn, U're gonna luv it hre
("Welcome to Valencia, Spain! Once the weather settles down, you're going to love it here")

The total words in the tweet are 14, with an AOpS of 6, or 43.0%. These are broken down into: 1 preposition: *2* (to); 2 adverbs: *dn* and *hre* (down and here); 1 verb: *luv* (love); and 2 contractions: *U're* and *gonna* (you are, going to).

Table 1 shows the AOpS of casual English in the twitter data, and **Fig. 1** shows a breakdown of distribution of casual English in terms of POS.

As indicated in **Table 1**, the AOpS of casual English items is, at 20.7%, perhaps surprisingly low. Regarding the POS findings, as shown in **Fig. 1**, only contractions stood out as being significantly more common. This is likely due to the fact that this category included both acronyms (*LOL, OMG* etc.) and non-standard contractions such as *imma*; however, we also included standard contractions where the writer had omitted the apostrophe, which were extremely frequent (*im, dont, wouldnt,* etc.). Therefore, the category was significantly broad.

Prepositions commonly included *4* or *2* (for and to). A large number of the counts for pronouns were for *u* (you), with a majority of conjunctions being variants of "and" and "because" (*an, n, coz, cuz* etc.). Counts for verbs were often *B* for "be," with a large number of gerunds (*in* or *in'* instead of writing the full *ing*, e.g., *dancin'*). Frequently occurring adverbs were variants of *sooo, nw, hre* (so, now, here) and adjectives were varied, although vowel lengthening for emphasis was common, e.g., *goooood*. Nouns were also varied, but a particularly common token was *ppl* (people). Interjections very frequently used vowel lengthening for emphasis, e.g., *aaaarrrgh*.

Table 1. Average words per sentence and occurrence per sentence (AOpS) of casual English.

Avg. words per sentence	Avg. casual English words per sentence
14.63	3.02 (20.7%)

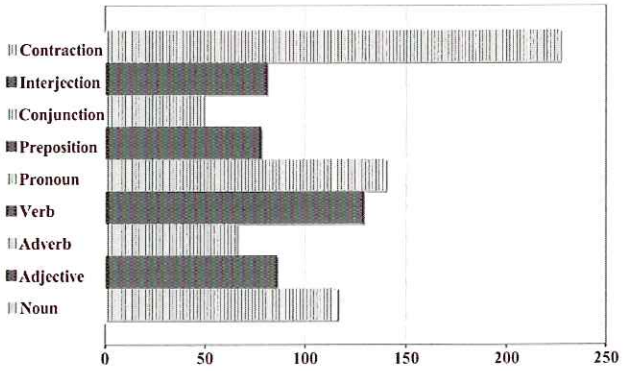

Fig. 1. Casual English distribution by POS.

Based on the discussion of various approaches in Section 2.1 and the results of the Twitter data analysis, we proposed to design a casual English generation system as follows. First, some superficial preprocessing such as lowercase conversion and URL detection/stripping will be conducted. As POS has shown to be of negligible influence in the analysis experiment, we decided not to use a POS tagger (e.g., a parser such as Enju[2]) on the input sentences to select targets for conversion, but to include common casual English contractions in a filter of "fixed words" taken from our previous normalization system's [1] token-to-token database, in order to tackle the problem of high occurrence of problematic contractions. This filter would also contain highly standard "slangifications" unlikely to be the target of varied phonetic conversion in real life, such as certain pronouns and interjections (e.g., "*u*" for "you," etc.).

For the conversion of remaining tokens, frequency per sentence was determined to be set at 21.0%, in line with the experiment results. Tokens selected randomly are split into phonemes using the CMU Pronouncing Dictionary. Finally, these phonemes are then converted using a manually compiled phoneme-to-phoneme database, which will be described in the following section.

2. http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

Table 2. Original alternative phoneme candidate database.

CMUDict Phoneme, Sound	CEGS Phoneme	CMUDict Phoneme, Sound	CEGS Phoneme
AA odd	*o,a*	K key	*k, kk*
AE at	*a*	L lee	*l, ll*
AH hut	*u,a*	M me	*m*
AO ought	*or, ar, aw*	N knee	*n, nn*
AW cow	*aw, ow, au*	NG ping	*n, n', nng, ngg*
AY hide	*y, ay*	OW oat	*ow, o, oa*
B be	*b, bb*	OY toy	*oi, oy*
CH cheese	*ch*	P pee	*p, pp*
D dee	*d, dd*	R read	*r*
DH thee	*v, th, d*	S sea	*s,$*
EH Ed	*e, eh*	SH she	*sh, shh*
ER hurt	*er, ur*	T tea	*t, tt*
EY ate	*a, ey*	TH theta	*t', th, f, ff*
F fee	*f, ff*	UH hood	*u, uh, oo*
G green	*g*	UW two	*oo, u*
HH he	*h, hh*	V vee	*v*
IH it	*i*	W we	*w, wh*
IY eat	*ee, y, i*	Y yield	*y*
JH gee	*g, j*	Z zee	*z, zz*
		ZH seizure	*j, jh, zh*

2.3. Phonetic Database

The database of candidates to represent phonemes using alternative spellings was constructed based on analysis of the large volume of casual English examples collected during the course of this research. The database consists of the 39 phonemes of the CMU Pronouncing Dictionary, with our alternative original phonemes as replacements. Most of the phonemes, although not all, have multiple replacement candidates which this method selects between randomly. For example, the word "everything" is split into the phonemes EH V R IY TH IH NG by CMUDict. In the database, the phoneme IY has multiple candidates of *ee, y*, and *i*, TH has multiple candidates of *t', th, f*, and *ff*, and NG has the multiple candidates *n, n', nng* and *ngg*. Thus, "everything" could be converted as various combinations, such as *evryffin', evreet'in, evrithingg*, etc.

The CEGS database is shown in **Table 2**. Capitals indicate the CMUDict phonemes, regular font exemplifies the sound as pronounced in an English word, and italics indicate the original CEGS alternative phoneme candidates. The CMUDict phoneme list and their sounds are reproduced from CMUDict homepage [6].

2.4. Phonetic Conversion

As described in Section 2.2, our analysis of casual token frequency found an average of 20.7% occurrence of casual English tokens per sentence. In the algorithm of CEGS, the proposed "Casual English Generation System," this is rounded up to 21.0% selection of input tokens

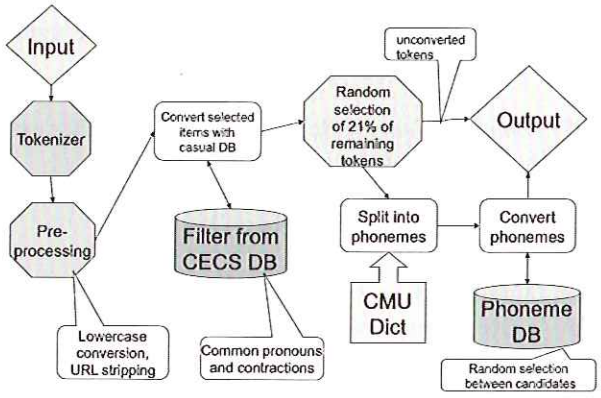

Fig. 2. Process of proposed system.

to be processed after the initial filtering stages. Although the secondary goal of the analysis experiment was to determine distribution of casual English tokens across POS categories, it was found that POS categories were not in fact shown to be a significant factor in the placing of casual English tokens overall. However, it was clear that certain words, particularly pronouns and some contractions, were very often written in the same way, e.g., "*u*" for "you," "*im*," for "I'm"; so these tokens were incorporated into CEGS using a filter consisting of a small section of our previous normalization system CECS' (Casual English Conversion System) [1] token-to-token database (with input and output reversed). An overview of CEGS is shown schematically in **Fig. 2**.

The process of CEGS is as follows. First, input (which is assumed to be regular English with no misspellings) is tokenized using a simple whitespace delimiter and removal of punctuation. Next, a single character string array of same length as the tokens in the input is created, in order to assign Boolean-type values of true ("process") or false ("do not process") to each token; this is because CEGS requires that only a minority of tokens are processed, as explained above.

The system then conducts some minor preprocessing on the input such as assigning "do not process" to certain tokens including URL or email indicators ("www," "http," "@," etc.). A second layer of preprocessing converts a fixed set of common standard tokens using the token-to-token database from the normalization system, CECS, which we proposed in a previous study [1].

After this stage, 21.0% of the remaining tokens are selected randomly using Python's random module,[3] which employs the Mersenne Twister as its generator. These tokens are assigned "process," while the rest are assigned otherwise. The processable tokens are split into their constituent phonemes using the CMU Pronouncing Dictionary, through the interface built into the Natural Language Toolkit [8]. Numbers signifying lexical stress and multiple outputs from the CMU Pronouncing Dictionary are removed, and the resulting phonemes are converted using our original phoneme database. Since many of these phonemes have multiple conversion candidates in the CEGS database, random selection between candidates is performed, giving different output each time in many cases. The output then consists of the sentence composed of filtered, processed and unprocessed tokens.

Thus, when an input sentence, for example the following:

Input: This is a regular English sentence which has been converted into casual English using the experimental system.

is processed by the system, the input is tokenized, lowercased, predetermined fixed tokens are flagged as non-objects of phonetic conversion, (in this sentence, only the indefinite article "a" is a fixed token) and 21.0% of the remaining tokens are randomly selected for phonetic conversion. In this example sentence, the selected tokens were: "this," "sentence," "which," "into," and "the," as seen in the output below.

Output: vi$ is a regular english sehnttuns wich has been converted intu casual english using da experimental system.

To illustrate the process at the token level, the word "sentence," when split by CMUDict, results in the phonemes "S," "EH," "N," "AH," "N," and "S." These are then converted into CEGS' original phoneme set, resulting in "sehnttuns" in this instance. Due to the random combinations of phoneme candidates as well as factors such as token length, some forms deviate from standard graphemes more than others (e.g., "sehnttuns" more than "wich"). The double inclusion of random selection, at both the token stage and phoneme stage, means that a second conversion of a sentence is almost certain to produce different results. A second processing of the above example sentence generated the following output:

Output (2nd time): this is a regular inglishh sentence which hazz bin converted intoo casual ingllish yoozzinng dda experimental system.

The selected tokens were mostly different: "English," "has," "been," "into," "English," "using," and "the"; and in the cases where the same tokens were selected as the first processing ("into," "the"), the graphemes are different, due to different combinations of phoneme candidates having been selected.

3. Evaluation Experiment

We conducted a Turing-type evaluation experiment on CEGS output using 50 human evaluators. In this section, we will describe the experiment method, present the results, and discuss the significance of our findings.

3.1. Experiment Method

Fifty human evaluators were questioned by an anonymous survey in our experiment. The main aim of our experiment was to perform a Turing-type test to ascertain whether human evaluators could distinguish CEGS output from human-authored sentences. The Turing test, a key concept in artificial intelligence commonly used to determine a machine's ability to demonstrate humanlike behavior, usually focuses on two-way natural language dialog between a machine and a human judge [3]. In this experiment, we did not use two-way communication, so the method was not a typical "Turing test"; but as the participants were judging the human-likeness of the machine-authored sentences in a blind environment, it is possible to define this method as a "Turing-type test."

The participants were shown 20 sentences, which consisted of 10 sentences converted from regular English by CEGS, and 10 sentences which were real life examples gathered from Choudhury's Twitter corpus. These sentences were mixed randomly. The evaluators were not told the purpose of the survey or that some sentences were "real" and some were "fake"; they were simply asked to evaluate each sentence on a semantic differential scale as follows:

1. *I am sure a machine made this.*
2. *This seems more machine-like than human-like.*
3. *I can't say either way.*

3. http://docs.python.org/library/random.html

4. *This seems more human-like than machine-like.*
5. *I am sure a human made this.*

Further, in order to determine the legibility of CEGS sentences in comparison to human ones, we also asked the evaluators to assess their comprehension of each sentence, again on a semantic differential scale:

1. *Don't understand at all.*
2. *Understand a little.*
3. *Understand somewhat.*
4. *Understand most.*
5. *Understand completely.*

Finally, we collected some background information on the evaluators in order to determine whether the following factors had any influence on their perception of CEGS output: age, gender, English ability, and the frequency of contact with social media-type English in their daily lives. The breakdown of the 50 participants was as follows. In the age category, 56% were aged 21–30, 24% were aged 31–40, 16% were aged 41–50, and 4% were aged 61 and over (no respondents fell into the categories of 10–20 or 51–60). By gender, 68% were female and 32% were male. In the category of English ability, 74% were native speakers, 12% were "non-native speakers with high confidence," 12% were "non-native speakers with medium confidence," and 2% were "non-native speakers with low confidence." Finally, in the category of contact with social media English, 22% answered "I see this kind of English every day," 32% answered "I see this kind of English sometimes," and 46% answered "I see this kind of English rarely" (no participants gave the answer "I never see this kind of English"). The CEGS sentences used in the experiment were taken from generic conversational sentences written in completely regular English which was then inputted into CEGS. Only the first output was taken (any second or further repeated outputs would be significantly different due to the random token selection), regardless of how legible the produced sentence was. An example of one of the experiment sentences produced by CEGS is shown below, followed by one of the "real" Twitter sentences also used in the experiment.

CEGS Sentence: it's always best 2 prepare before u do something like that. u just wa$tt time otherwise. don't b a fool.
(Input: *It's always best to prepare before you do something like that. You just waste time otherwise. Don't be a fool.*)

Human Sentence: Do u knw wat hurtz d most? Itz wen u had made sum1 feel Special yesterday & d same person 2day claiming u 2 b da most Unwanted person...

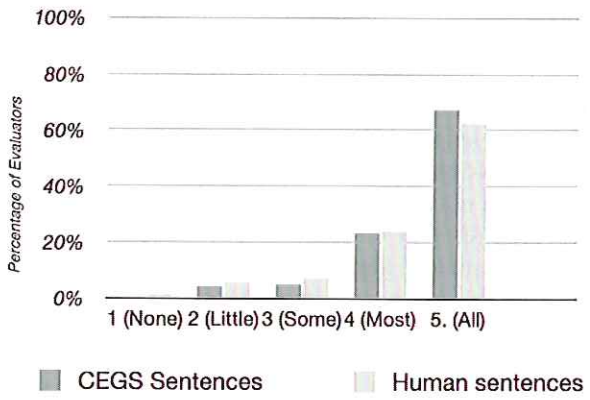

Fig. 3. Comparative reader comprehension of CEGS and human-authored sentences.

(Translation: *Do you know what hurts most? It's when you had made someone feel special yesterday and the same person today (is) claiming you to be the most unwanted person...*)

3.2. Experiment Results

The results for the comparative legibility of the sentences will be followed by the results of the Turing-type test according to the categories of the participants.

The legibility of the CEGS sentences and human sentences is compared in **Fig. 3**. The numbers 1–5 indicate the semantic differential scale of "Don't understand at all" (1) to "Understand completely" (5), as previously described in Section 3.1. Thus, higher numbers signify higher reader comprehension.

A detailed breakdown of the evaluators' average "human-likeness" ratings for the CEGS and human sentences is shown in **Table 3**. The scores are given as an average of evaluators' rankings for all 20 sentences, on the semantic differential scale of 1–5 where 1 is "I am sure this sentence was made by a machine" and 5 is "I am sure this sentence was made by a human." Thus, higher numbers signify higher human-likeness. Highest and lowest scores are indicated in bold.

3.3. Discussion

The key findings of our evaluation experiment, as shown in the previous section, were that: a) although human-likeness scores were on average higher for human-authored sentences, the gap with CEGS is not extremely large; b) understanding of sentences on average was high, and broadly similar between CEGS and human sentences, with CEGS actually showing slightly higher understanding overall; this indicates that human creativity may be more difficult to understand than "artificial creativity"; and c) somewhat surprisingly, although CEGS output cannot always pass for human in the Turing test, human-written sentences cannot always pass for human either. Filtering of responses by evaluator category produced some unexpected results, which we will discuss below.

Table 3. Human-likeness evaluation of both CEGS and human-authored sentences by category.

Evaluators by category	CEGS Sentences	Human Sentences
All Evaluators	3.1*	3.7
Gender		
Female	3.19	3.8
Male	3.03	3.6
English Ability		
Native Speakers	3.18	3.8
Non-native, high confidence	**2.77**	3.6
Non-native, mid confidence	3.18	3.5
Non-native, low confidence	**3.6**	**2.9**
Social Media English Contact		
Every day	3	3.9
Sometimes	3.14	3.8
Rarely	3.2	3.6
Age		
Age 21–30	3.06	3.8
Age 31–40	2.96	3.7
Age 41–50	3.56	3.6
Age 61+	3.5	**4**

*(SD score of 1–5, where 5 is certainty of human authorship)

We will also discuss some sentences which effectively "fooled" the respondents and suggest why this may have occurred.

In the breakdown of responses by category, we found that regarding age, the gap between the human-likeness assessment of CEGS and human sentences after the age of 41 (0.04 and 0.5 SD points respectively), was smaller than the gap before the age of 41 (0.74 and 0.74 SD points); in other words, users over the age of 41 were more likely to be "fooled" by CEGS. We attribute this tendency to an unfamiliarity with social media English among older participants. We also found that male evaluators were slightly more likely to score both CEGS and human-authored sentences as being "machine-like," whereas females were more inclined to believe that the sentences were written by humans.

Although we had expected that native speakers of English would be the most capable of correctly detecting artificial casual English, in fact the group that ranked CEGS sentences with the lowest human-likeness scores were the non-native speakers with high confidence, who gave CEGS sentences an average score of 2.77 in contrast to 3.18 from native speakers. We speculate that this may be due to the fact that the highly able non-native speakers could observe patterns in English analytically from their experience of learning English consciously, whereas native speakers' ability is not consciously learned and thus perhaps less rational. The mid and low confident non-native speakers had smaller differences between the scores of CEGS and human-authored sentences (0.32 and 0.7 SD points respectively) than the highly confident native speakers (0.83 SD points), which was in line with our expectations. The low-confidence category actually evaluated CEGS sentences as being more human-like than human ones, which should indicate success for CEGS; however, the sample was very small.

We had assumed that familiarity with social media English would make evaluators harder to "fool" with CEGS sentences, and this was shown to be the case, as human-likeness scores for CEGS were the lowest for those who answered "I see this kind of English every day," and highest for those who answered "I see this kind of English rarely." The reverse was true with human-written sentences: higher exposure to social media English made evaluators more likely to assess human-written sentences as human, and lower exposure made evaluators more like to give slightly lower human-likeness scores to human-written sentences.

Although the average scores for all 20 sentences showed that CEGS sentences could not quite achieve parity with genuine human-authored sentences, at 3.1 to 3.7 respectively, some individual sentences countered this overall trend. For example, the two CEGS sentences shown below both scored an average of 3.5 from all 50 evaluators; which puts them closer to the "human-like" end of the semantic differential scale than the "machine-like" end, and is a higher score than 40% of the actual human-authored sentences.

CEGS Sentence: da thing is, i could never support a person like that. no matter wut they did 2 redeem themselves, da act izz arllredee done.
(Input: *The thing is, I could never support a person like that. No matter what they did to redeem themselves, the act is already done.*)

CEGS Sentence: thi$ is getting really stupid. i dont care what peepull r saying or what people think. i just want 2 live my life how i want. come on.
(Input: *This is getting really stupid. I don't care what people are saying or what people think. I just want to live my life how I want. Come on.*)

The second sentence appeared to be particularly convincing, with 60% of the evaluators ranking it as 4 ("This seems more human-like than machine-like") or 5 ("I am sure a human made this"). This is an example of CEGS' random selection resulting in infrequent and reasonably believable conversions such as "peepull."

However, spellings generated by CEGS are not usually as credible as this. The sentence below was the lowest ranked of all 20 sentences, with an average score of 2.72.

CEGS Sentence: haha i dont know about any of that. i said what i thought azz soon as i sedd it. maybe my mouth izz too quick for my bbreyn?
(Input: *Ha ha, I don't know about any of that. I said what I thought as soon as I said it. Maybe my mouth is too quick for my brain?*)

In this case, the random token selection converted more tokens than usual, and the spellings were not particularly

natural, e.g., "sedd" and "bbreyn." This was remarked on by one of the evaluators in the free comment space at the end of the survey, who observed: "*When the word is spelt oddly but not abbreviated, it seems more machine-like (so 'bbreyn' for instance).*"

In contrast, it is difficult to pinpoint why the human-authored sentence with the lowest average score, 3.0, was not deemed to be very human-like:

Human Sentence: Only vry rarely has a person 2 the same extent as Obama captured the world's attention & given its ppl hope 4 a better futur.
(Translation: *Only very rarely has a person to the same extent as Obama captured the world's attention and given its people hope for a better future.*)

The abbreviated forms seen in this sentence, such as "vry," "ppl" and "4" are quite common in casual written English, so it is not clear why this sentence was ranked as less human-like than 70% of CEGS sentences. We surmise that the content of the sentence was perhaps somehow deemed to be artificial; as an evaluator commented: "I think the content of these messages probably slightly influenced whether I thought they were human or machine, more than the way they were written..." Since our evaluation experiment did not give any clue to the evaluators regarding what we were actually testing, some may have assumed that the content itself was artificially generated (whereas the sentence in question was in fact taken from Twitter).

4. Conclusion

We have described CEGS, a system for generating casual English short sentences from regular English input using a phonetic database approach. With the results of an evaluation experiment, a Turing-style test using fifty human evaluators, we have shown that although CEGS has not yet reached a level of naturalness completely equal to human-authored sentences, the gap between the two is not significantly large, and that some CEGS sentences actually outperformed several of the human sentences in human-likeness scores. The CEGS sentences' average score of 3.1 on the 5-point scale, where 3 indicated "*I can't say either way*" whether the sentences were human-authored or machine-authored, can be considered a pass of the Turing test in the classical sense; the standard condition for passing the Turing test is when a human cannot reliably distinguish that a machine is a machine. This suggests that CEGS' potential for use in automatically generating Twitter or SMS-style slang is high.

In future, we aim to achieve greater human-likeness in CEGS output, through modifications such as adding or altering phoneme candidates in the database or implementing some "intelligent selection" to the currently random token selection and phoneme candidate selection, perhaps via optimization techniques such as genetic algorithms or by supervised learning methods guided by human user judgments. This would enable the creation of more natural forms. Another improvement to human-likeness would be the incorporation of wider types of casual English, e.g., non-dictionary slang and acronyms, in the token-to-token filter from the token-to-token normalization system database. However, multiple candidates, random selection or weighting should be used in order to preserve variety in output of such expressions.

References:

[1] E. Clark and K. Araki, "Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English," Proc. of the 12th Conf. of the Pacific Association of Computational Linguistics, Kuala Lumpur, Malaysia, 2011.

[2] A. Ritter, C. Cherry, and B. Dolan, "Unsupervised modeling of Twitter Conversations," Proc. of HLT-NAACL 2010, Los Angeles, California, pp. 172-180, 2010.

[3] A. Turing, "Computing Machinery and Intelligence," Mind, Vol.59, No.236, pp. 433-460, 1950.

[4] A. Aw, M. Zhang, J. Xiao, and J. Su, "A phrase-based statistical model for SMS text normalization," Proc. of the COLING/ACL 2006 Main Conf. Poster Sessions, pp. 33-40, 2006.

[5] C. Kobus, F. Yvon, and G. Damnati, "Normalizing SMS: are two metaphors better than one?," Proc. of the 22nd Int. Conf. on Computational Linguistics, pp. 441-448, 2008.

[6] http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Accessed on 2012.7.30)

[7] M. D. Choudhury, Y. R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the sampling strategy impact the discovery of information diffusion in social media?," Proc. of the 4th Int. Conf. on Weblogs and Social Media, Washington DC, USA, 2010.

[8] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.

Name:
Eleanor Clark

Affiliation:
Language Media Laboratory, Division of Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University

Address:
Kita 14, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

Brief Biographical History:
2006 Received Bachelor's degree from the Faculty of Languages and Cultures, School of Oriental and African Studies, University of London
2009 Received Master's degree from Graduate School of Letters, Hokkaido University
2009- Doctoral Student, Graduate School of Information Science and Technology, Hokkaido University

Main Works:
• E. Clark and K. Araki, "Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English," Elsevier Procedia in Social and Behavioral Sciences, Vol.27, pp. 2-11, 2011.
• E. Clark and K. Araki, "Two Database Resources for Processing Social Media English Text," Proc. of LREC 2012, pp. 3790-3793, 2012.

Name:
Kenji Araki

Affiliation:
Language Media Laboratory, Division of Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University

Address:
Kita 14, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

Brief Biographical History:
1988 Received Ph.D. in Electronic Engineering, Hokkaido University
1988-1991 Researcher and Lecturer, Hokkai Gakuen University
1991-1998 Associate Professor, Hokkai Gakuen University
1998-2002 Associate Professor, Hokkaido University
2002- Professor, Hokkaido University

Main Works:
- K. Araki and K. Tochinai, "Acquisition Words by Inductive Learning and Recognition Words Using Certainty," Trans. of the Institute of Electronics, Information and Communication Engineers D-II, Vol.J75-D-II, No.7, pp. 1213-1221, 1992.
- K. Araki and Y. Momouchi, "Concept Learning from Japanese Copular Sentences Using Heuristics," Proc. of the Australian Joint Conf. on Artificial Intelligence, pp. 466-473, 1994.
- K. Araki and K. Tochinai, "Effectiveness of Natural Language Processing Method Using Inductive Learning," Proc. of IASTED-AISC, pp. 295-300, 2001.

Membership in Academic Societies:
- American Association of Artificial Intelligence (AAAI)
- The Institute of Electrical and Electronics Engineers (IEEE)
- Information Processing Society of Japan (IPSJ)
- The Japanese Cognitive Science Society (JCSS)
- The Institute of Electronics, Information and Communication Engineers (IEICE)