3P1-IOS-2a-2

A Phonetic Rule-based Approach for Generating Casual English Sentences

Eleanor Clark[*1] and Kenji Araki[*1]

[*1] Graduate School of Information Science and Technology, Hokkaido University

We present a system for generating casual English short sentences from regular English input using a phonetic rule-based approach. This is addressed as an AI task, with the potential application of generating Twitter-style sentences for marketing or other communication purposes. Our aim was to automatically produce sentences that would appear to a human reader to be indistinguishable to sentences which are the result of human creativity. To evaluate the performance of the system, we conducted Turing-type tests with human readers, to consider firstly "human-likeness", and also "legibility" of the sentences. In this paper, we discuss the overall design of the system, the custom-made phoneme database, and the process and results of our evaluation experiment.

1. Introduction

The proliferation of highly irregular casual written English on in electronic communications including emails, chat applications, SMS (Short Message Service), microblogs such as Twitter, has created a large volume of publicly available data, but the irregularity and creativity of the language poses a problem for NLP (Natural Language Processing) applications such as machine translation, information extraction, ontology creation, and summarization [Clark A. 2003, Ritter et al. 2010]. 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. We attempted 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 our method, we utilized a phoneme-by-phoneme approach, which attempts to mimic SMS (short message service) or Twitter-type phonetic spellings by selecting replacement candidates at the phonemic level. Selected tokens are split into phonemes using the CMU Pronouncing Dictionary[1], and these phonemes are then converted into the multiple alternative phonemes in our database. As this method can produce highly creative phonetic slang, it is necessary to strike a balance between "interesting" and "difficult to understand". 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.

2. CEGS: A Casual English Generation System

2.1 System Overview

One important point in casual English sentence design is that, usually, not all tokens (words) in a given sentence are irregular. Even if only a small proportion of tokens per sentence consists of casual English items, this is often enough to render the sentence incomprehensible to a non-native speaker or to a machine translation application, as we have shown in previous research on social media English [Clark and Araki, 2011]. Thus, frequency of casual English tokens per sentence was selected based on prior linguistic analysis of 320 tweets, in which casual English items and their POS were manually tagged, in order for the method to reflect the human creation of casual English sentences in a more natural way. Our analysis found an average of 21.67% occurrence of casual English tokens per sentence. In the algorithm of CEGS, a "Casual English Generation System", this is rounded up to 22% selection of input tokens 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, we found that POS categories were not in fact shown to be a significant factor in the placing of casual English tokens overall. However, we found 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 are incorporated into CEGS using a filter consisting of a small section of the token-to-token database from our previous research focusing on normalization of casual English text (with input and output reversed) [Clark and Araki, 2011]. An overview of CEGS is shown schematically in Figure 1.

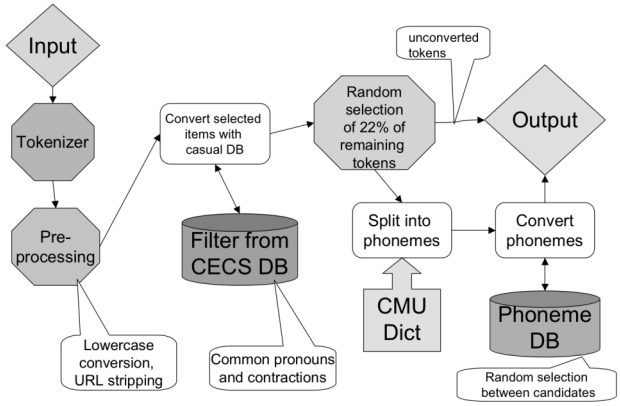

Figure 1: System flow of CEGS.

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. We then conduct some minor

Contact: Eleanor Clark, Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 8, Kita-ku, Sapporo, Japan 060-0814, Tel: +81-11-706-7389, Fax: +81-11-709-6277, eleanor@media.eng.hokudai.ac.jp

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 our previous research.

After this stage, 22% of the remaining tokens are selected randomly using Python's random module, which employs the Mersenne Twister as its generator[2]. 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 [Bird et al., 2009]. 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.

2.2 Phonetic Database

The alternative phoneme representation is constructed based on analysis of the large volume of casual English examples collected during our research. The database consists of the 39 phonemes of the CMU Pronouncing Dictionary, with our alternative original phonemes as replacements. These phonemes were selected based on their occurrence in casual English words in the Twitter corpus used in our research [Choudhury et. al 2010]. Most of the phonemes, although not all, have multiple replacement candidates, which our method selects between randomly. For example, the word "everything" is split into the phonemes EH V R IY TH IH NG by CMU Dict. In our 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.

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 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
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, 68.1% were female and 31.9% 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 completely 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.)*

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

[2] http://docs.python.org/library/random.html

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.. (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

We present the results for the comparative legibility of the sentences, followed by the results of the Turing-type test in detail according to the categories of the participants.

The legibility of the CEGS sentences and human sentences is compared in Fig. 2. The numbers 1-5 indicate the semantic differential scale of "Don't understand at all" (1) to "Understand completely" (5). Thus, higher numbers signify higher reader comprehension.

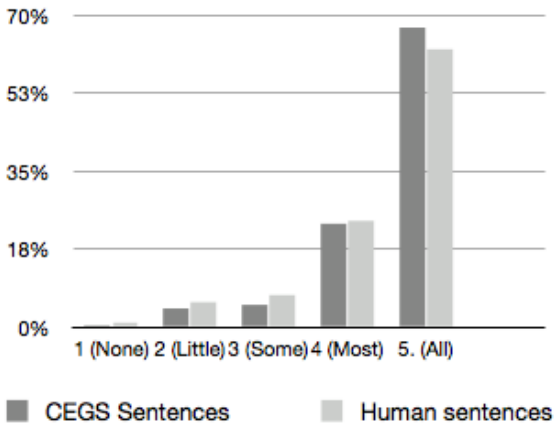

Figure 2: Comparative reader comprehension of CEGS and Human-authored sentences

A detailed breakdown of the evaluators' average "human-likeness" ratings for the CEGS and human sentences is shown in Table 1. 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.

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**

Table 1: Human-likeness evaluation of both CEGS and human-authored sentences by category (SD score of 1-5 where 5 is certainty of human authorship)

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. 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 the gap between the human-likeness assessment of CEGS and human sentences decreased as the age of the evaluators increased, which we attribute 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 less confident non-native speakers had smaller differences between the scores of CEGS and human-authored sentences, which was in line with our expectations.

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 thing. 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 humanlike:

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. (*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. Conclusions

We have described CEGS, a system for generating casual English short sentences from regular English input using a phonetic rule-based approach. With the results of our first 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. 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.

References

[Bird et al. 2009] Bird, S; Klein, E. and Loper, E., *Natural Language Processing with Python*. O'Reilly Media, 2009.

[Choudhury et al. 2010] Choudhury, M. D.; Lin, Y.R.; Sundaram, H.; Candan, K, S.; Xie, L. and A. Kelliher. How does the sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th International Conference on Weblogs and Social Media*, Washington DC, USA, May 2010.

[Clark. A, 2003] Pre-processing very noisy text. In *Proceedings of Workshop on Shallow Processing of Large Corpora,* Lancaster, UK, March 2003, pp. 12–22.

[Clark. E and Araki 2011] Clark, E and Araki, K. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. In *Proceedings of the Twelfth Conference of the Pacific Association of Computational Linguistics*. Kuala Lumpur, Malaysia, July 2011.

[Ritter et al. 2010] A. Ritter, C. Cherry and B. Dolan. Unsupervised modeling of Twitter Conversations. In *Proceedings of HLT-NAACL 2010*, Los Angeles, California, June 2010, pp. 172–180.