FluTCHA: Using Fluency to Distinguish Humans from Computers

Kotaro Hara, Mohammad T. Hajiaghayi, Benjamin B. Bederson
Computer Science Department
University of Maryland, College Park
{kotaro, hajiagha, bederson}@cs.umd.edu

ABSTRACT

Improvements in image understanding technologies are making it possible for computers to pass traditional CAPTCHA tests with high probability. This suggests the need for new kinds of tasks that are easy to accomplish for humans but remain difficult for computers. In this paper, we introduce Fluency CAPTCHA (FluTCHA), a novel method to distinguish humans from computers using the fact that humans are better than machines at improving the fluency of sentences. We propose a way to let users work on FluTCHA tests and simultaneously complete useful linguistic tasks. Evaluation demonstrates the feasibility of using FluTCHA to distinguish humans from computers.

Categories and Subject Descriptors

H5.m. Information interfaces and presentation.

General Terms

Human Factors; Languages

Keywords

CAPTCHA; Human computation

1. Introduction

ReCAPTCHA [1,2] asks people to read and enter characters from an image attempts to: 1) distinguish humans from computers; and 2) harness human power to transcribe text from an image. Although it is often used in web registration to filter out spambots, recent improvements in computer vision technologies make it possible to break ReCAPTCHA tests with nearly 100% accuracy [7]. This suggests that we need to consider new tasks that are more difficult for computers to solve, while remaining easy for humans.

Yamamoto *et al.* previously introduced a new type of CAPTCHA, SS-CAPTCHA, that focuses on the difficulty of assessing fluency of a sentence [6]. That work provides users with a number of fluent sentences and a set of machine-translated sentences that are not fluent. The users are asked to identify the sentences that are fluent. Since there is no effective method of automatically evaluating fluency of a sentence, it is hard for computers to identify the correct answers but it is easy for humans. The test is, however, vulnerable against dictionary attacks because the system uses sentences collected from public sources from the Internet for *both* poor quality *and* fluent sentences. Thus once spammers know where the source sentences are taken from, they would know which sentences are written by humans.

In this paper, we introduce a prototype CAPTCHA system called FluTCHA that asks users to perform linguistic tasks such as *editing* non-fluent machine-translated texts to make them fluent. The task is easy for native language speakers, but remains difficult for computers [5]. FluTCHA is safe from dictionary attacks because it uses only translated sentences, and relies on editing instead of multiple choice questions. This, moreover, has an advantage that the FluTCHA process results in linguistic data that could be useful to improving machine translation systems.

Original Japanese	大統領が署名を事実上拒否し、交渉が行き詰まっていることが理由という(The negotiations stalled because the president refused to sign.)
Machine Translation	Effectively refused to sign the President, that is why negotiations stalled.

Table 1: A non-fluent machine translated sentence is generated from a sentence in a Japanese newspaper article by Google Translate. The author translated the original Japanese sentence into the English sentence in parentheses.

2. Prototype FluTCHA System

FluTCHA collects Japanese sentences from news web sites and automatically translates them into English using Google Translate (example shown in Table 1). Many translated sentences are understandable by humans even when they are not fluent.

FluTCHA consists of two types of human tasks: paraphrasing and grading (Figure 1). In a paraphrasing task, a person being tested (*answerer)* edits a translated sentence. The FluTCHA interface shows a translated text, highlighting a sequence of consecutive words. The interface also provides the same sentence where a textbox substitutes the highlighted area; the answerer must type in a paraphrased and more fluent text into the textbox. Note that, while being tested, the answerer also improves the quality of translation (*i.e.,* known as "post-editing"). Once the answerer finishes paraphrasing, the original translated sentence and the paraphrased sentence are sent to the next step.

The fluency of the original translation and the edited sentence are evaluated in a second human activity of grading (by *graders*). FluTCHA shows the original translation and the paraphrased sentence in random order. Graders are asked to grade the fluency of each sentence on a nine-point scale. This enables FluTCHA to evaluate whether the answerer is a human or a bot by evaluating how much the paraphrased sentence's fluency improved compared to the translation. Graders are also asked to grade the similarity in meaning between the two sentences, which allows us to notice if an answerer entered a phrase that changes the meaning of the original translation. Thus, FluTCHA generates a corpus of human edits of machine-translated sentences while testing for humanness. These edited sentences could, for example, be used in machine translation systems.

While FluTCHA does not return test results instantaneously, paraphrased sentences can be graded and results sent back to answerers fairly quickly as shown below. To make the system even faster, we could use the retainer approach introduced by

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742759

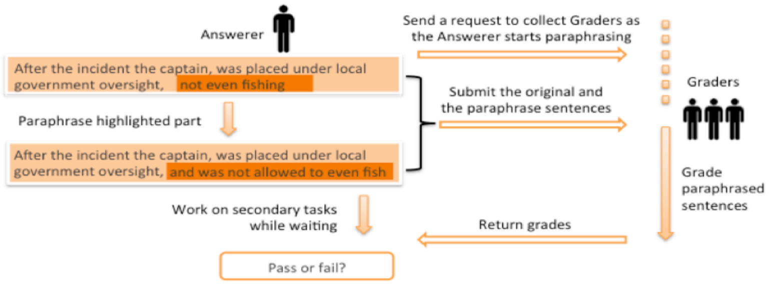
Answerer
After the incident the captain, was placed under local government oversight, not even fishing
Paraphrase highlighted part
After the incident the captain, was placed under local government oversight, and was not allowed to even fish
Work on secondary tasks while waiting
Pass or fail?
Send a request to collect Graders as the Answerer starts paraphrasing
Submit the original and the paraphrase sentences
Return grades
Graders
Grade paraphrased sentences

Figure 1. The FluTCHA system involves two groups of humans: answerers and graders. FluTCHA sends a paraphrase along with the original sentence to graders, and sends back the result once graders evaluate the fluency of the original (machine-translated) sentence and the paraphrase sentence.

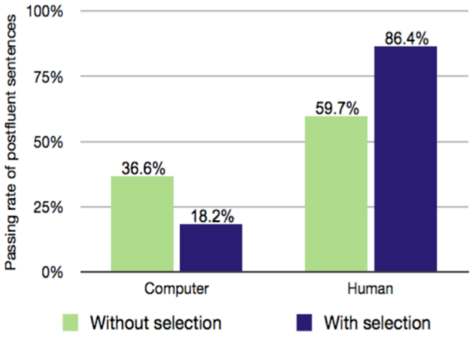

Figure 2 The figure shows that computer success rate drops from 36.6% to 18.2% and human success rate increases from 59.7% to 86.4% as we select suitable test sentences for paraphrasing tasks.

Bernstein *et al.* [4] so that there are graders available when needed.

3. Method

We evaluated whether FluTCHA is capable of distinguishing humans from computers.

Data: We collected 478 sentences from 44 news articles from a Japanese news website. We then translated them into English using Google Translate.

Answerers: We asked five native English speakers to volunteer for a study.

Highlighting: First, answerers highlighted parts of the machine-translated sentences that are not fluent. This was done to create sentences for the evaluation paraphrasing tasks.

Paraphrasing: We asked answerers to paraphrase the portions of the sentences that were highlighted by other participants to increase their fluency.

Pivot: To simulate a bot's activity, we used Google Translate to translate the highlighted parts of the translated sentences from English to Japanese, then from Japanese back to English, which resulted in disfluent sentences. We used these disfluent sentences to compare against human paraphrasing.

Grading Task: We posted tasks on the Amazon Mechanical Turk (mTurk) online labor market, and recruited workers to grade the fluency of the translated, paraphrased, and pivoted sentences along with similarity in meaning. We asked workers on mTurk to grade two sets of sentences per assignment for 5 cents. We collected 898 grades from 83 distinct workers. Each set of sentences was graded from 6 to 17 times. We considered tasks as passing if any given sentence showed at least 1 point improvement compared to the original translation.

4. Result

Paraphrased sentences generated by humans passed single tests at a rate of 59.7% (536/898), while pivoted sentences generated by computers passed at a rate of 36.6% (329/898) (Figure 2). The paraphrasing task took 39.2 sec to complete on average (we did not measure the time for highlighting since it is not a part of the FluTCHA task). The somewhat poor performance was in part due to the fact that some sentences were difficult for humans to improve (*e.g.,* because an original sentence was already fluent).

To assess the potential of FluTCHA, we selected sentences that were hard for computers and easy for humans. We wanted to know how well FluTCHA could work if we identified sentences that are suited for the task a priori. To select those sentences, we used an F-measure to balance precision and recall [3]. We selected sentences where FluTCHA separated human from computer with an F-measure above 0.8. We defer the discussion of how to select "good" sentences in real-system to future work.

The result is shown in Figure 2. The success rate for humans increased from 59.7% to 86.4% while the success rate for computer work decreased from 36.6% to 18.2%. By calculating cumulative probabilities, we estimate that people can pass the test 98.1% of the time after 2 trials. The expected number of trials for humans to pass the test is 1.16 (1/86.4%) and the expected number of trials for computers to pass the test is 5.49 (1/18.2%).

5. Discussion

We have shown that FluTCHA is capable of distinguishing humans from computers with a success rate of 36.6% and 59.7% by computers and humans, respectively. We have also shown that by selecting sentences that are easy for humans to improve, these figures improve to 18.2% and 86.4%. This begs the question of whether this is good enough; while the success rate of computers is a concern, the current performance of FluTCHA has a ready path towards improvement. And existing CAPTCHA systems already have fairly poor performance. As we collect a larger set of sentences, we could curate sentences for a test that are highly suited for separating humans from computers. For curation, we could ask an answerer to paraphrase two sentences; the first sentence would be used to separate a human from a program, and the second sentence is there for our benefit to evaluate if it is suitable for the test. We should also evaluate system performance with a population with varying fluency levels in the future.

6. REFERENCES

1. Ahn, L. Von, Blum, M., Hopper, N.J., and Langford, J. CAPTCHA : Using Hard AI Problems For Security. *Science*, .
2. Ahn, L. Von, Maurer, B., Mcmillen, C., Abraham, D., and Blum, M. reCAPTCHA : Human-Based Character. *JCCFS*, September (2008), 1465–1468.
3. Baeza-Yates, R., Ribeiro-Neto, B., and others. *Modern information retrieval.* ACM press New York, 1999.
4. Bernstein, M.S., Brandt, J., Miller, R.C., and Karger, D.R. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. *UIST 2011*, 33–42.
5. Callison-Burch, C., Bannard, C., and Schroeder, J. Improved statistical translation through editing. *EAMT-2004*, (2004).
6. Yamamoto, T., Tygar, J.D., and Nishigaki, M. CAPTCHA Using Strangeness in Machine Translation. *ICAINA2010*.
7. Yan, J. and Ahmad, A.S. El. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms. *ACSAC 2007*.