Technical Disclosure Commons

Defensive Publications Series

September 01, 2016

CAPTCHA using Word Relationships

Jason Fedor

Aaron Malenfant

Marco Zennaro

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation

Fedor, Jason; Malenfant, Aaron; and Zennaro, Marco, "CAPTCHA using Word Relationships", Technical Disclosure Commons, (September 01, 2016)
http://www.tdcommons.org/dpubs_series/266

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

CAPTCHA using Word Relationships

ABSTRACT

CAPTCHAs are challenge questions employed by websites to prevent access by automated programs (bots) while allowing access to human users. The effectiveness of a CAPTCHA in distinguishing between humans and bots is sometimes reduced, because as bots get more powerful, they are able to solve certain types of challenge questions almost as well as humans. This disclosure describes CAPTCHAs based upon word relationships. Solving the CAPTCHAs requires demonstrating natural language understanding. Automated programs are not yet comparable to humans in their ability to understand natural language. CAPTCHAs of this disclosure provide useful techniques to distinguish human users from bots.

KEYWORDS

- CAPTCHA
- Challenge question
- Natural language
- Word relationships
- Visually impaired

BACKGROUND

Owners of web resources often require that only humans access a given web resource. For example, restricting access only to humans can ensure that the web resource cannot be attacked, e.g., by automatically launching large numbers of requests for access. Further, restricting access only to humans can also prevent automated extraction of content by crawling

of a web resource. Techniques are in use that permit a web resource to forbid access by bots — programs that simulate human activity.

For example, to distinguish humans from bots, the web resource poses a challenge question to the entity requesting access. A challenge question is easy for a human to solve but difficult for a computer. Such challenge questions are known as CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart). CAPTCHAs should be capable of being auto-generated by a system providing the web resource (e.g., a computer).

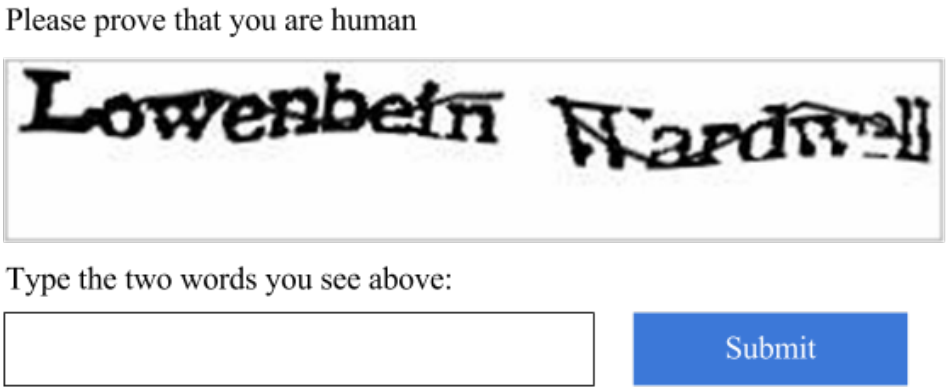


Fig.1: An example CAPTCHA

A user who wants access to a web resource is first presented with a challenge in the form of a CAPTCHA as shown in the example of Fig. 1. The distorted words `Lowenbein' and `Wardwell' are presented to the entity requesting access to the web resource. Humans are good at recognizing visual patterns, even when those are deliberately distorted in form or otherwise modified (e.g., by additional wavy lines). Thus, humans can easily recognize the words in the CAPTCHA and solve the challenge. A bot that relies on image processing and optical character recognition techniques however cannot pass the challenge as easily.

There are at least two problems with current CAPTCHAs, such as the example in Fig. 1.

1. Bots are constantly becoming smarter, and are already somewhat capable of correctly recognizing words that are visually distorted or otherwise modified.
2. CAPTCHAs such as the example shown in Fig. 1 are not easily solvable by visually impaired humans. An alternative challenge presented to the visually impaired is an audio challenge that requires the user to listen to a garbled voice and identify numbers or words that are being spoken. Advances in speech recognition now allow bots to more easily pass such audio CAPTCHAs.

DESCRIPTION

This disclosure describes CAPTCHAs that are based on the relationships of words to one another. For example, the challenge question requires the user to categorize certain words into classes, ask the user to identify the relationship between two words, ask if the entity represented by one word is part of the entity represented by another word etc. This type of challenge requires complex understanding of natural language that computers are worse at than humans. Although the types of challenge questions described in this disclosure based on word relationships are difficult for bots to solve, the questions themselves are computer-generable. The challenge questions are computer-generable via algorithms that have capabilities of natural-language understanding and access to a large and spanning corpus of words. Notwithstanding that the challenge questions are computer-generable, bots cannot easily answer such challenge questions.

Examples of use

This section describes examples of CAPTCHAs based upon categorization of words, pose “part of a whole” relationship questions, or require a user to determine the relationship

between two words. CAPTCHAs that rely on other types of word relationships can also be generated similarly.

Example 1

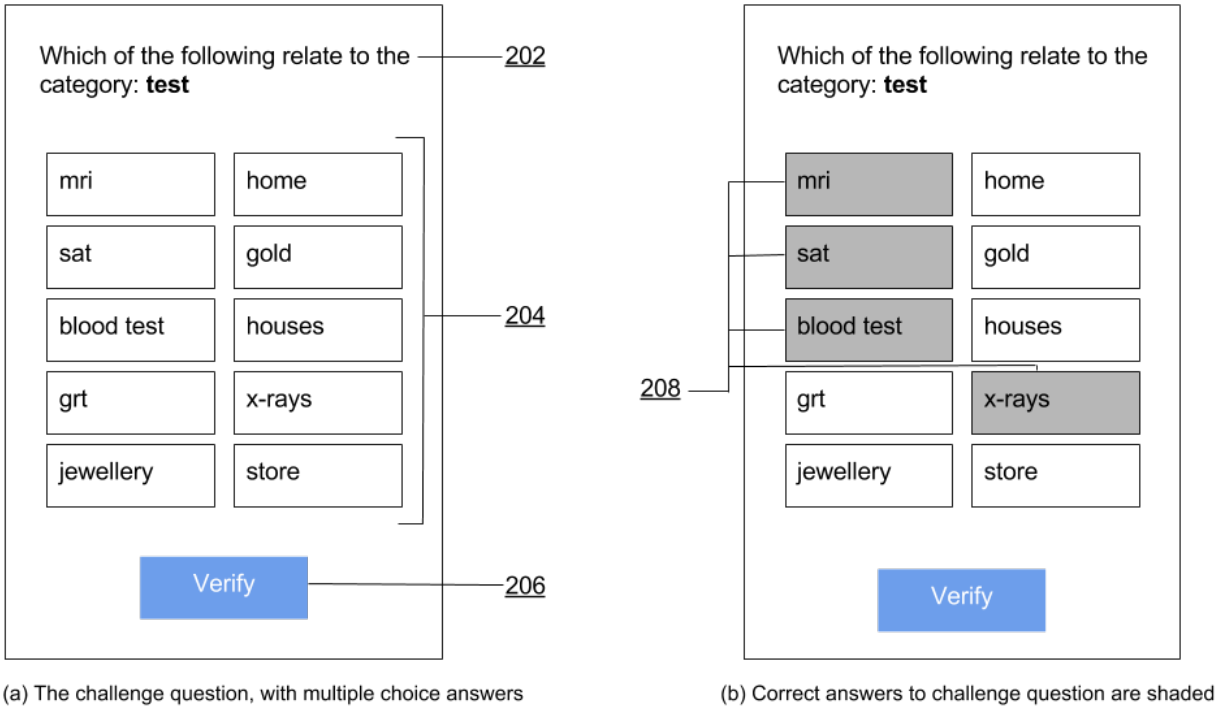


Fig. 2: An example CAPTCHA based on categorization of words

Fig. 2 shows an example user interface for a CAPTCHA based upon categorization of words. Fig. 2(a) shows a challenge question (202), which requires the user to select out of multiple choices (204) words related to the category "test". The user interface permits selection of words (e.g., by tapping or clicking on a word) and submission of selected words using the verify button (206). For example, an English-speaking human user can easily identify that "sat" is indeed a type of (scholastic) "test" and that "mri", "blood test" and "x-rays" are each a type of (medical) "test". Accordingly, a human user can select these as the answers, as indicated by

shaded boxes (208) in Fig. 2(b). However, a bot cannot identify the relationship between the category in the question and each of the words available as potential answers because the bot does not have sufficient natural language capabilities to identify the relationships between the category in the question and each of the words available as potential answers.

Example 2

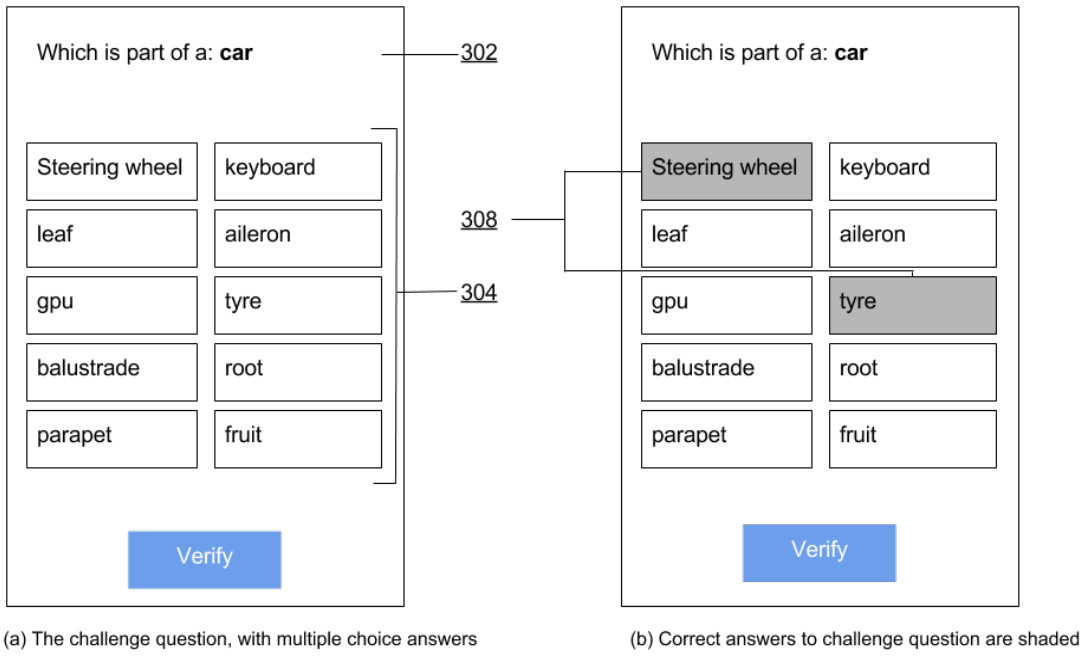
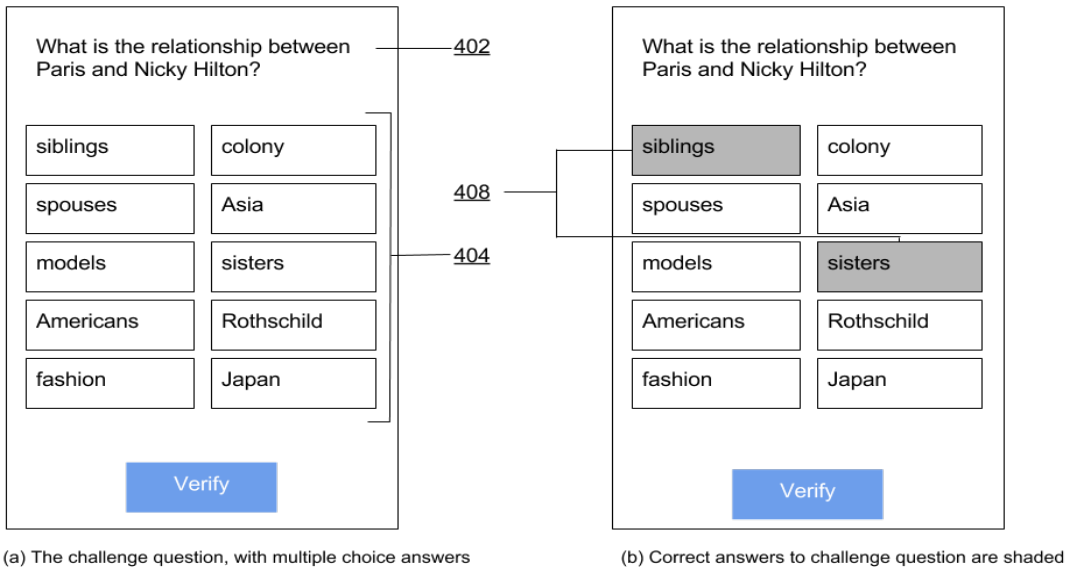


Fig. 3: An example CAPTCHA based on `which is part of which'

Fig. 3 shows an example of a CAPTCHA that requires a user to select from presented multiple choices (304) words that are associated with items that are parts of an entity described by a word in the challenge question (302). In the example of Fig. 3(a), the entity in the challenge question is “car.” Of all the possible answers provided, an English-speaking human user can easily identify that only “steering wheel” and “tyre” are part of “car”, and accordingly, select these words, as indicated by the shaded boxes (308) in Fig. 3(b). However, a bot cannot identify

answers that are associated with items that are parts of the entity described by the challenge word because the bot does not have sufficient natural language capabilities.

Example 3



(a) The challenge question, with multiple choice answers

(b) Correct answers to challenge question are shaded

Fig. 4: An example CAPTCHA based on relationship between two entities

Fig. 4 shows an example of a CAPTCHA that requires the user to determine the relationship between two words within the challenge question (402). In the example of Fig. 4(a), the two words are "Paris" and "Nicky Hilton." Of the answers provided (404), an English-speaking human user can easily identify that the correct answers are "siblings" and "sisters", and accordingly, select these words, as indicated by the shaded boxes (408) in Fig. 4(b).

Generation of CAPTCHAs

A database of word relationships is populated by collecting language data from web crawlers and various other sources. In one example, the word relationship is in the form of category membership. As shown in Fig. 5, category "presidents" includes "Truman", "Kennedy",

"Obama" etc.; category "legalese" includes "But not limited to", "Inter alia", "Sub judice" etc.; and category "fruit" includes "melon", "orange", "apple" etc.

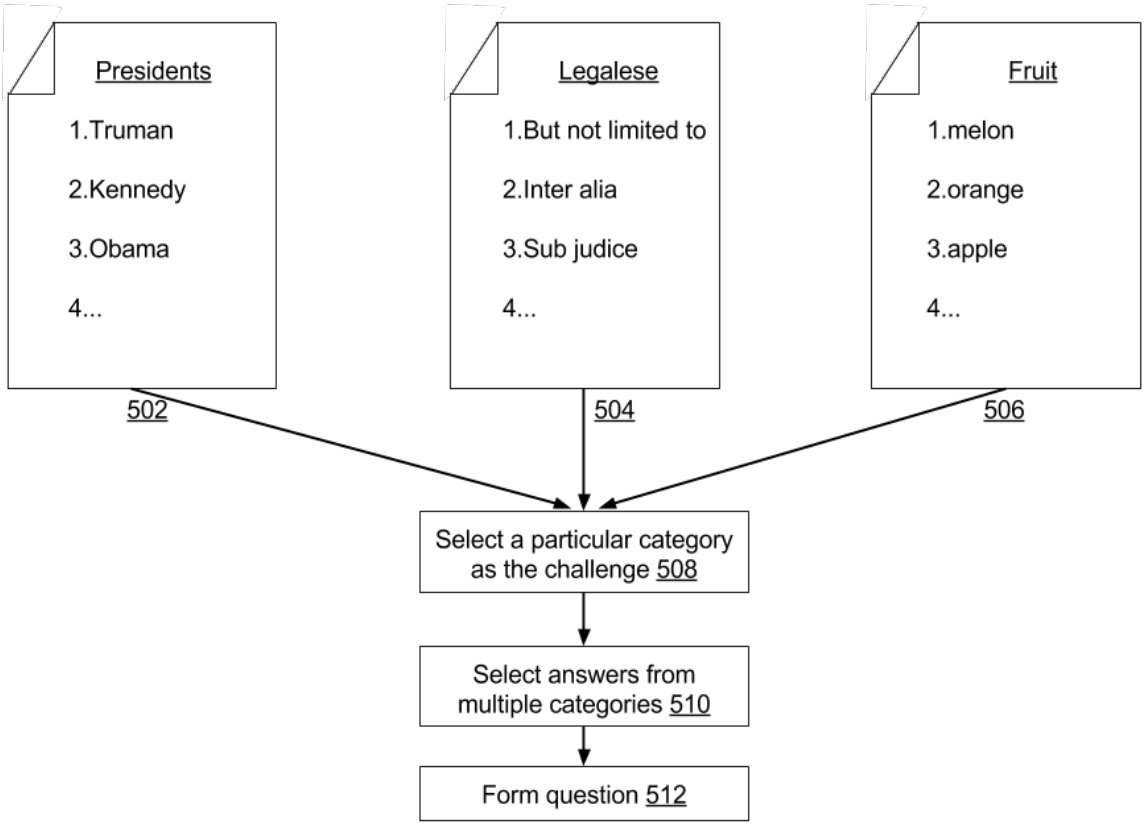


Fig. 5: Formation of questions from word relationships

Fig. 5 shows an example of how challenge questions are formed using word relationships. Out of several categories (502-506), a particular category is selected as the challenge (508). Items from multiple categories, including the particular category, are selected to form possible answers to the challenge (510). The challenge question is formed (512) using selected category as question and selected items as possible answers. In the example of Fig. 5, an example challenge question is "Which of the following relate to the category *legalese*?", and potential answers include "melon", "But not limited to", "Truman", "Inter alia", "Kennedy" and "apple". The correct answers are "But not limited to" and "Inter alia".

Suitability for visually impaired users

Traditional CAPTCHAs, such as the example shown in Fig. 1, rely on visual features (e.g., wave lines) that make it difficult for bots to recognize text. Such CAPTCHAs are unsuitable for visually impaired users. The CAPTCHAs of this disclosure rely on conceptual relationships between words. While Figs. 2-4 show examples of user interfaces visually, the CAPTCHAs can easily be adapted for visually impaired users. For example, the CAPTCHA question can be narrated via audio as "Which of the following words describe the relationship between Paris and Nicky Hilton?" followed by narration of the possible answers, e.g., "siblings, spouses, colony, Japan, sisters". Visually impaired users can provide their answers, e.g., by speaking the words "siblings" or "sisters".

Learning using user-generated answers

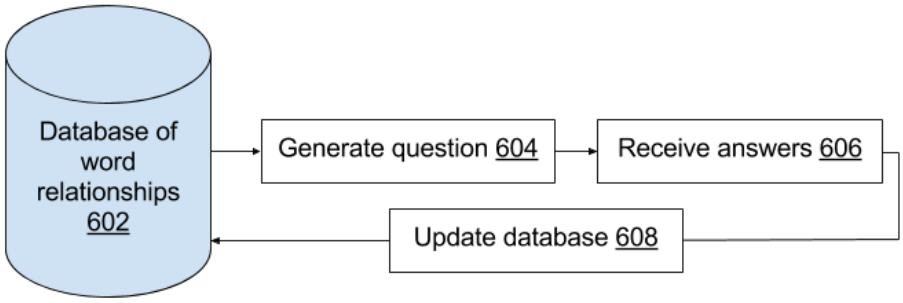


Fig. 6: Updating database using user-generated answers

The technique described here can also present CAPTCHAs that include challenge questions where not all answers are known with certainty. Answers selected by a user can then be used to improve the quality of the database of word relationships, as shown in Fig. 6. From a database of word relationships (602), a question is generated (604). Answers (606) are received,

which indicate membership of certain words in certain categories. Using information from the answers, the database is updated (608).

CONCLUSION

This disclosure describes CAPTCHAs based upon the relationship of words to other words in a corpus. Such CAPTCHAs contain challenge questions that are computer-generable via algorithms that have some capabilities of natural-language understanding and access to a large and spanning corpus of words. Although the challenge questions themselves are computer generable, bots do not have sufficient natural language capabilities to easily answer such challenge questions. CAPTCHAs as described herein can easily be presented in visual or audio format, thus providing accessibility respectively to both the hearing and/or the visually impaired. They can also be presented in any natural language.