©2015 IEEE. This is the submitted and accepted version of the paper.
The final revised version will appear in the proceedings and the digital library of IEEE.

Automatic Generation of Authentication Questions from Private Messages

Ming Li
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501 Japan

Keishi Tajima
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501 Japan
Email: tajima@i.kyoto-u.ac.jp

***Abstract*—In this paper, we propose a method for automatically generating authentication questions in social network services (SNSs) and mail account services. When a malicious user obtains a password of some SNS or mail account, the malicious user can access private messages posted/sent to or from the account, and also messages posted in closed SNS groups the account participates in. In order to prevent it, many systems pose additional questions when a suspicious user tries to login to an account or try to access messages in a closed group. Our method automatically generates such authentication questions for an account or group by using the messages in that account or group. Our method shows one of the messages with substituting one noun with a blank, and ask the accessing user what word was there. To detect fake users, we need to select a noun that is sufficiently difficult for fake users to infer based on general knowledge and information on the Web. We select a noun based on two factors. First, for each candidate noun, we compute its co-occurrence degrees on the Web with other words in the same message. If a noun has high co-occurrence degrees with other words in the message, the noun is probably easy for fake users to infer. Second, our system collects coordinate terms (co-hyponyms) of each candidate noun, and calculate the same co-occurrence degrees of them. If there are coordinate terms that have higher co-occurrence degrees than a candidate noun, we expect that the noun is difficult for fake users to infer because those coordinate terms seem to them more likely to be the answer. We developed four methods of noun selection based on these two factors. Our preliminary experiment shows that the former factor produces more difficult questions than the latter, but it often produces questions that are too difficult even for authentic users.**

***Keywords*-social network service; SNS; mail account; knowledge-based authentication; KBA**

I. Introduction

On social network services (SNSs), such as Facebook, users can make their accounts either open or closed. Messages posted to or from a closed account can be read only by the account owner and the friends. On some SNSs, users can also create closed groups, where only the group members can post or read messages. Mail account services also have similar access control; only the owner of an account can read messages sent to or from the account.

In addition, to prevent illegal access by malicious users who have stolen the passwords of some accounts, many services monitor the access patterns of users, such as IP addresses and devices used for the access, and when they detect some anomaly, they require the accessing user to provide additional information for proving the authenticity, such as the answer to a pre-registered authentication question.

Many users, however, use the same question-answer pairs for many services, and the malicious user who has stolen the password from other services may also obtain the question-answer pair. In order to reduce such a risk, users should use a different question for each service, but users rarely do it. There are also users who do not register their questions.

In this paper, we propose a method of automatic generation of authentication questions that are specific to each SNS/mail account or each SNS group. The idea is to use private messages of that account or group. When a suspicious user tries to login to an account, our system selects a private message posted/sent to or from that account, show it with substituting one word in it with a blank, and ask the user what word was there in the message. Because private messages are only known to the owner of the account and the friends, we expect that fake users cannot answer it while the authentic user can. Similarly, when a suspicious user tries to access messages in a closed group, our system generates a question from one of the messages posted in that group.

In this approach, it is important to select and remove an appropriate word. If it can be easily inferred from other words in the message only based on general knowledge and information on the Web, even fake users can answer the question. For example, a question:

> It's our kid's birthday! I baked a _____!

is too easy. Even fake users can infer the answer "cake" from the words "birthday" and "bake".

Our method selects a noun in the messages based on two factors. The first factor is the co-occurrence degrees of the noun and other words within the same message. We compute the co-occurrence degree of two words by using the hit counts of Web search engines. That is, we use the co-occurrence degrees on the Web. If a noun has high co-occurrence degrees on the Web with some words in the same message, even fake users can easily infer the noun based on the general knowledge or information on the Web.

The second factor is the co-occurrence degrees of coordinate terms (co-hyponyms) of the noun. Given a candidate noun in a message, we collect coordinate terms of the noun, and compute their co-occurrence degrees with the other words in that message. If there are many coordinate terms that have higher co-occurrence degrees than the candidate noun, we expect that the noun is difficult to infer for fake users because these coordinate terms seem to fake users more likely to be the answer. For example, the question shown above is not easy if the answer is "scone". Even though the word "scone" also has a high co-occurrence degree with "bake", its coordinate terms

"cake" and "cookie" have even higher co-occurrence degrees with "bake", and seem more likely to be the answer.

In this paper, we propose four methods that select a noun based on these two factors. The result of our preliminary experiment shows that the former factor produces more difficult questions but it often produces questions that are too difficult even for authentic users.

II. Related Work

Authentication by a question on the knowledge that the user is supposed to have is called knowledge-based authentication (KBA), and has been widely used. Many systems use pre-registered questions, which are called static KBA, but there are also systems that dynamically produce questions, e.g., based on the past access history of the user. Our method is also an example of such dynamic KBA.

Toomin et al. [1] has proposed an authentication scheme for SNS groups based on the knowledge shared by the group members, such as "the name of the dog of Susan". In their scheme, however, the questions are manually specified by the users. That is, their scheme is static KBA.

There has also been studies on dynamic KBA for SNSs. Gampa et al. [2] proposed a method that uses the preferences of a user. When a user receives a friend request from an account which claims to be his/her real-life friend, the system automatically generates a question on the preferences of the user receiving the request. This method is based on the assumption that people must know well the preferences of their real-life friends. This method, however, cannot prevent illegal access to some accounts by a real-life friend of the account owner. It cannot also be applied to closed groups whose members are not necessarily real-life friends.

Yardi et al. [3] proposed an authentication scheme for closed groups. They show a group member's photo posted in the closed group, and require the accessing user to answer the name of the member in the photo. However, if some other photos of the member are publicly available elsewhere, even a fake user can answer it by finding such photos, e.g., by some similarity image retrieval techniques. In order to prevent such attacks, Polakis et al. [4] proposed a method that selects photos for which face recognition programs fail, and also developed a technique to distort face images in a photo so that image matching programs do not work while the friends are still able to identity the person.

Our method also has the problem of the inference attack based on publicly available knowledge. In order to prevent it, we select the word to remove based on its co-occurrence degrees on the Web. A disadvantage of the photo-based method is that we cannot use it when no photo of the members has been posted. On the other hand, our method uses text messages, and it is unlikely that there has been no message. Another disadvantage of the photo-based method is that the variation of answers is bound by the number of members, while the answers to our questions have wider variety.

Yet another disadvantage of the photo-based method is that the photo which is not public is shown to suspicious users. Our method has the same problem. We need some method of selecting a message that does not reveal too sensitive information. It is an important issue for future research.

In language learning, a question generated from a sentence by removing a word in it is called a cloze test. There have been some studies on automatic generation of cloze tests for computer-aided learning [5], [6]. Their criteria for word selection are, however, completely different from ours.

There have been several proposals of methods for collecting coordinate terms of a given word. We use the method proposed by Ohshima et al. [7]. Their method collects coordinate terms of a word t by searching the Web with two queries "t and" and "and t", and extracting phrases that appear both immediately after "t and" and immediately before "and t".

III. Proposed Method

In this section, we explain the details of our proposed method of question generation.

A. Selecting Nouns as Initial Candidates

First, our method assigns part-of-speech tags to words in the given messages by using GoSen[1], a morphological analyzer for Japanese language. We only use nouns for creating questions because they are most strongly related to the informational contents of messages, and therefore most useful for distinguishing authentic users who know the informational contents of messages and fake users who do not know them.

On the other hand, verbs and adjectives are often substitutable with some similar words unless they appear as parts of frequent phrases. When they are substitutable, questions produced by removing them are difficult even for the authentic users, and when they are parts of frequent phrases, questions produced by removing them are easy even for fake users.

Some nouns have the same problem. Nouns representing general concepts are often substitutable with other words unless they appear as parts of frequent phrases. In order to eliminate such nouns from candidates, we further classify nouns. GoSen classifies nouns in Japanese language further into 14 categories. We only use nouns in the following three categories: general nouns, proper nouns, and "nouns convertible to adverbs" category. The last category mainly includes nouns representing some time period, such as October. We do not use nouns in the other categories, such as "nouns convertible to verbs", which only include nouns representing general concepts, such as competition and departure. Note that general noun category in the classification by GoSen does not include these nouns, and mainly include names of things.

GoSen assigns a part-of-speech tag to each individual word, but sentences often include compound nouns. When there is a contiguous sequence of nouns, we aggregate the nouns into a compound noun. In Japanese, this simple rule suffices. In English, we should use parse trees to detect compound nouns. If a compound noun includes at least one noun classified into one of the three categories explained above, we include that compound noun as a candidate for question generation.

[1]lucene-gosen at GitHub, https://github.com/lucene-gosen/lucene-gosen

B. First Factor: Co-occurrence with Other Words

When generating a question, we need to select a word which is difficult to infer for fake users. In order to select such a word, we compute co-occurrence degrees of each candidate noun with other words in the same message.

There are many ways to define co-occurrence degrees of two words. They can be classified into two types: symmetric ones and asymmetric ones. Our purpose of calculating co-occurrence degrees is to measure how easy for fake users to infer the removed word from another word. The inference in the opposite direction, i.e., from the removed word to another word, is not important at all. For this reason, we define an asymmetric co-occurrence degree of t_1 with t_2, denoted by $C(t_1|t_2)$, by the formula below:

$$C(t_1|t_2) = \frac{|D(t_1) \cap D(t_2)|}{|D(t_2)|}$$

where $D(t) = \{d \in U | t \text{ appears in } d\}$. U is the corpus we use for measuring co-occurrence degrees, and we use the Web as the corpus as explained before.

Given a candidate noun t_q in a message m, we compute $C(t_q|t)$ for each of the other words t in m. We approximate $|D(t)|$ by the hit count of the query "t" given by a Web search engine. $|D(t_1) \cap D(t_2)|$ is also approximated by the hit count of the query "t_1 t_2".

$C(t_q|t)$ can be regarded as an approximation of the probability that a fake user infer t_q by associating t with it based on general or public knowledge. It can also be regarded as an approximation of the probability that a fake user selects t_q as the best choice among the words with which the user can associate t (with ignoring the constant factor for normalization, i.e., for making the sum of all probabilities be 1).

We then approximate the probability that a fake user can infer the correct answer t_q from all the other words in the message m altogether, which is denoted by $P(t_q|m)$. Note that we ignore the constant factor for normalization. First, let us discuss the approximation by the formula below:

$$P_{all}(t_q|m) = 1 - \prod_{t \in words(m), t \neq t_q} (1 - C(t_q|t))$$

where $words(m)$ is the set of words in the message m. This formula computes the probability that a fake user infer t_q from at least one of the other words in m. This method, however, does not reflect the real inference process by users. This method computes the probability that a user infer one most likely noun from each of the other words, try these answers one by one, and none of them is the correct answer. A user is, however, allowed to answer only once.

Next, let us discuss the approximation by the formula below:

$$P_{max}(t_q|m) = \max_{t \in words(m), t \neq t_q} C(t_q|t).$$

This method is based on the assumption that a fake user selects one word in m which seems the most useful clue and infer the answer from that word. Users, however, sometimes infer the answer from more than one words. In order to include such cases into consideration, we should generalize t in this formula to subsets of words. First, we generalize t_2 in the definition of $C(t_1|t_2)$ to a set of words T_2 by the formula below:

$$C(t_1|T_2) = \frac{|D(t_1) \cap \bigcap_{t_2 \in T_2} D(t_2)|}{|\bigcap_{t_2 \in T_2} D(t_2)|}.$$

$C(t_1|T_2)$ can be regarded as an approximation of the probability that a fake user selects t_q as the best choice among the words with which the user associate a set of words T_2. We then generalize t in the definition of $P_{max}(t_q|m)$ to a set of words T by the formula below:

$$P_{ideal}(t_q|m) = \max_{T \subseteq words(m) \setminus \{t_q\}} C(t_q|T).$$

This method, however, requires an exponential number of queries to a search engine, and is infeasible when we have 20 or more words in a message. For this reason, we consider two methods of further approximating it, and compare them in our experiment. One method is to approximate it by $P_{max}(t_q|m)$, i.e., we assume that the value of $C(t_q|T)$ is dominated by $t \in T$ that has the largest $C(t_q|t)$ value. Another method is to use $P_{all}(t_q|m)$, in other words, we approximate the best inference from a subset of $words(m)$ by the disjunction of the inferences from each word in $words(m)$.

In summary, we compare two approximation methods $P_{all}(t_q|m)$ and $P_{max}(t_q|m)$ in our experiment.

C. Second Factor: Coordinate Terms

Another factor we use for selecting a noun is the co-occurrence degrees of the coordinate terms of the candidate noun. We select a candidate noun that has coordinate terms that have higher co-occurrence degrees, as explained before.

For each candidate noun, we collect its coordinate terms by using the method proposed by Ohshima et al. [7], as explained in Section II. Let $CO(t_q)$ denote the set of obtained coordinate terms of t_q. We then compute $P(t|m)$ for each $t \in CO(t_q)$, and compute the score of a candidate noun t_q, denoted by $S(t_q)$, by the formula below:

$$S(t_q) = \sum_{t \in CO(t_q), P(t|m) > P(t_q|m)} P(t|m)$$

That is, we sum up $P(t|m)$ of coordinated terms that seem to fake users more likely to be the answer than t_q.

For example, suppose t_q has three coordinate terms t_1, t_2, t_3. If $P(t_q|m) = 0.3$, $P(t_1|m) = 0.4$, $P(t_2|m) = 0.4$, $P(t_3|m) = 0.1$, then $S(t_q) = 0.8$. On the other hand, even if $P(t_q|m)$ has a smaller value 0.25, if $P(t_1|m) = P(t_2|m) = P(t_3|m) = 0.2$, i.e., if none of the coordinate terms seem more likely to be the answer, $S(t_q) = 0$.

We select a noun t_q with the highest score $S(t_q)$. To compute $P(t|m)$, we use either $P_{all}(t|m)$ or $P_{max}(t|m)$.

D. Noun Selection Methods

Based on these two factors, we designed the following four methods of selecting a noun:

1) select a noun t_q with the lowest $P_{all}(t_q|m)$,

2) select a noun t_q with the lowest $P_{max}(t_q|m)$,
3) compute $S(t_q|m)$ by using $P_{all}(t_q|m)$, and select a noun t_q with the highest $S(t_q|m)$,
4) compute $S(t_q|m)$ by using $P_{max}(t_q|m)$, and select a noun t_q with the highest $S(t_q|m)$.

We compared these four methods in our experiment.

E. Other Additional Procedures

When we apply our method to a user accessing a closed group, we first select messages that are either posted or responded by the accessing user, and generate a question by using one of these messages. It is because each member of a closed group may not have read all the messages in the group.

When the correct answer is a compound noun, and the answer by the user is its subsequence, our method shows the message "please describe it in more detail" and ask the user to revise the answer instead of simply denying the access.

IV. Experiment

We conducted a preliminary experiment for evaluating the validity of our basic approach. We collected 80 messages from a closed group in Facebook, of which one of the author is a member, and selected nouns by using our four methods.

The result of the method 1 and the method 2 was the same, and the words with underlines in the following sentences were selected as the top-three candidates.

$Q1$) Added photos to Tokyo Branch End-of-Year Party 2011
$Q2$) Prof. YY talked in FM. Didn't understand from the very beginning. "Sampled-data control theory" ...(omitted)
$Q3$) Wow, already uploaded those in the morning during the event lunch break.

All these words seem difficult for fake users to infer from other words in the sentences, but all of them are quite rare words ("lunch break" is not rare, but "event lunch break" is an unusual phrase), and even the group members may not be able to correctly remember them. On the other hand, the method 3 selected the following words as the top-three candidates:

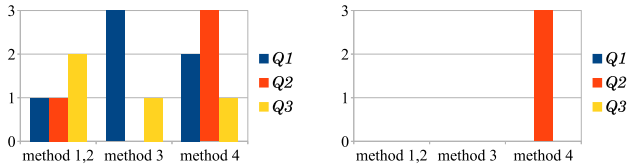
$Q1$) Added photos to Tokyo Branch End-of-Year Party 2011
$Q2$) Prof. YY talked in FM. Didn't understand from the very beginning. "Sampled-data control theory" ...(omitted)
$Q3$) Branch director, MG, let us know which day is convenient for you.

and the method 4 selected the following three:

$Q1$) Added photos to Tokyo Branch End-of-Year Party 2011
$Q2$) (omitted) ...but it shows his personality, and it was fun.
$Q3$) MC, thank you for taking pictures.

Because "event lunch break" is an unusual phrase and our method found no coordinate term of it, it was not selected by the method 3 and 4. Instead, the method 3 and 4 selected more general words, "branch director," "personality," and "taking pictures" (which is a compound noun in Japanese).

We hired 18 volunteers, assigned 6 volunteers to each of the method 1 (or 2, which produced the same questions), 3, and 4. We then ask them to answer top three questions generated by the corresponding method. Three of the six assigned to each method read all the 80 messages before answering questions, and the other three did not read the messages. Figure 1 shows how many of them could give a correct answer. Note that Q_1 is the same question in all methods, but the results are different because we assigned different six users to each method.



Users who read messages Users who did not read messages

Figure 1. The number of users who gave the correct answer.

V. Discussion and Conclusion

In this paper, we proposed four methods that generate authentication questions by removing one noun in a private message, and compared them by a preliminary experiment.

The method 1 and 2 tend to select very rare nouns, such as long compound nouns, and as a result, even users who have read the messages often cannot give the correct answer. The method 3 and 4 tend to select more general nouns because they select a noun only when it has more popular coordinate terms. As a result, even users who have not read the messages gave the correct answer to Q_2 of the method 4, where the answer is "personality", which is a very general word.

These observations suggest that we can improve our methods by eliminating too long compound nouns, and/or by selecting words that have either no coordinate term or many more popular coordinate terms. We are planning experiments of these improved methods with a bigger data set.

The method explained in this paper has one component that is specific to Japanese language: the classification of nouns. We are also planning experiments with other languages.

References

[1] M. Toomim, X. Zhang, J. Fogarty, and J. A. Landay, "Access control by testing for shared knowledge," in *Proc. of CHI*, 2008, pp. 193–196.
[2] N. K. Gampa, R. A. Khot, and K. Srinathan, "Let only the right one IN: privacy management scheme for social network," in *Proc. of the 5th International Conference on Information Systems Security*, ser. LNCS, vol. 5905, 2009, pp. 310–317.
[3] S. Yardi, N. Feamster, and A. Bruckman, "Photo-based authentication using social networks," in *Proc. of the First Workshop on Online Social Networks*. ACM, 2008, pp. 55–60.
[4] I. Polakis, P. Ilia, F. Maggi, M. Lancini, G. Kontaxis, S. Zanero, S. Ioannidis, and A. D. Keromytis, "Faces in the distorting mirror: Revisiting photo-based social authentication," in *Proc. of CCS*, 2014, pp. 501–512.
[5] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proc. of HLT*. ACL, 2005, pp. 819–826.
[6] M. Agarwal and P. Mannem, "Automatic gap-fill question generation from text books," in *Proc. of the Workshop on IUNLPBEA*. ACL, 2011, pp. 56–64.
[7] H. Ohshima, S. Oyama, and K. Tanaka, "Searching coordinate terms with their context from the web," in *Proc. of WISE*, 2006, pp. 40–47.