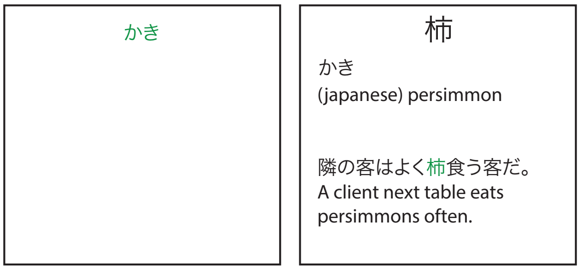
Kotonoha: An Example Sentence Based Spaced Repetition System

Arseny Tolmachev Sadao Kurohashi
Kyoto University

arseny@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

1 Introduction

Word learning is one of the most time consuming parts of language learning, especially when mastering languages to advanced level. For example, N1, the most difficult level of Japanese Language Proficiency Test requires knowing about 10000 words. Furthermore, learners need not only know the words *independently*, they also need how words are used *in context*: together with other words.



Question for the reading card
User is expected to remember
the writing and meaning.

A flashcard for the word 柿
(answer card)

Figure 1: Flashcards for the word "柿"

Word learning is often done using flashcards – a way of organizing information into question-answer pairs. An example of a flashcard for the word "柿" is shown on the Figure 1. An answer card has several fields: kanji writing, reading, definition and reference example sentence. A question card usually has only one of those fields. A student is expected to remember remaining information. We denote a flashcard that has only the specified field for a question as a <*field*> card. For example, we call the card on the Figure 1 a reading card because the question contains only reading field.

Flashcard systems frequently use Spaced Repetition technique to optimize learning process. The technique is based on observation that people tend to remember things more effectively if they study in short periods spread over time (*spaced repetition practice*) opposed to *massed practice* (i.e. cramming) [5, 2]. Anki[1] is probably a most widely known open source Space Repetition System (SRS). One of drawbacks is that a learner has to create all flashcards to learn. There exist decks created by community of various quality, however they are not as effective as manually created ones. The more serious problem is lack of context in flashcards. Cards usually have the form like displayed on Figure 1. Namely, the question card has no context information. Even if there is an example sentence present, it is static (it does not change from repetition to repetition) and can not show the full spectrum of word usage.

We have developed a Kotonoha[2] Spaced Repetition system which uses automatically extracted example sentences to form flashcard questions. Its sources are available on github[3]. Kotonoha shows a new example sentence on each flashcard repetition. Also, it makes adding and keeping flashcard deck process easy by requiring learners to input only writing information of words. Kotonoha adds lexical information like meaning, pronunciations and glosses from dictionary automatically. Example sentences are attached without any learner intervention as well.

Kotonoha is a useful system for learning Japanese language. However, it is also designed to be a very useful source of research data. For example, Kotonoha makes it possible to evaluate effect of example sentences on learning words of a foreign language. Usually, methods of evaluating quality of example sentences are often subjective (researchers ask learners of their opinion) or time consuming (need to perform tests on large number of participants). In contrast to that, a spaced repetition system can be used to directly measure whether example sentences are useful for remembering words or not without frequent manual intervention from researchers.

[1] http://ankisrs.net
[2] https://kotonoha.ws
[3] https://github.com/kotonoha/server

2 Spaced Repetition

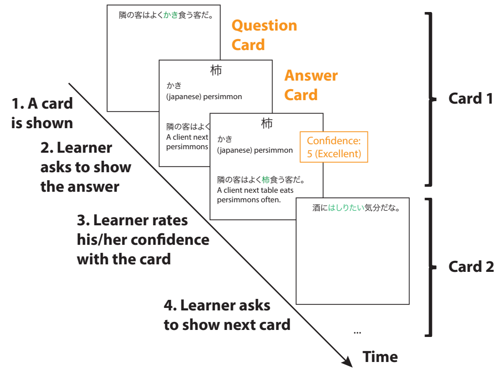

隣の客はよくかき食う客だ。
Question Card
柿
かき
(japanese) persimmon
隣の客はよく
A client next
persimmons
Answer Card
柿
かき
(japanese) persimmon
Confidence: 5 (Excellent)
隣の客はよく柿食う客だ。
A client next table eats persimmons often.
酒にはしりたい気分だな。
...
Card 1
Card 2
1. A card is shown
2. Learner asks to show the answer
3. Learner rates his/her confidence with the card
4. Learner asks to show next card
Time

Figure 2: Learning process with flashcards

Spaced repetition is based on *spacing effect*. It is an phenomenon first documented by Ebbinghaus in 1885. People tend to remember things more effectively in long term if they use short study periods which are spread over time. Moreover, it was shown that if spacing between practices increases, people remember things even better [3].

Learning process using spaced repetition is done in short sessions. In each session a student reviews several flashcards. A flashcard can be shown to user from two sources: either it is a new one, so it was never reviewed before, or it is an old one, so it was reviewed before. Old cards are shown only after a certain period of time passes from the previous repetition. The period of time is computed based on card repetition history and algorithm-dependent. For a brief review of scheduling algorithms for spaced repetition systems please refer to the paper [6].

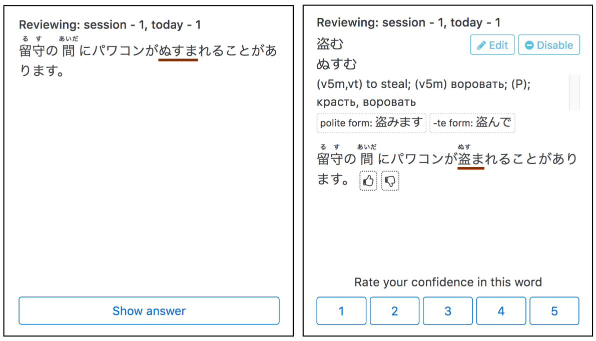
A repetition process sketch for a single flashcard is shown on the Figure 2. The process consists of three repeated steps. In each step, a student:

1. Views a *question* card
2. Tries to remember the answer
3. Views card's *answer*
4. Evaluates own confidence in the answer
5. Proceeds to the next card

After the repetition, each card is scheduled based on the confidence score reported by learner. Cards with high confidence are scheduled after intervals which are exponentially increasing. In contrast, cards with low confidence are scheduled after a short interval.

3 Kotonoha SRS

Kotonoha is a web-based open-source Spaced Repetition System. It's main distinctive feature is that it uses example sentences for reading and writing question cards. Kotonoha uses algorithm for Spaced Repetition from SuperMemo family [9] similarly to Anki.

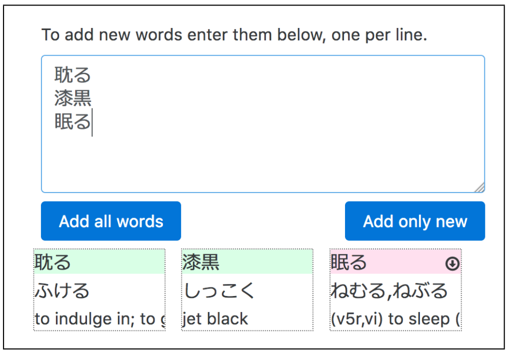

Reviewing: session - 1, today - 1
留守の間にパワコンがぬすまれることがあります。

Show answer

Reviewing: session - 1, today - 1
盗む
ぬすむ
(v5m,vt) to steal; (v5m) воровать; (P); красть, воровать
polite form: 盗みます
-te form: 盗んで
留守の間にパワコンが盗まれることがあります。
Rate your confidence in this word

Figure 3: Screenshot of reading card for the word "盗む" in Kotonoha SRS

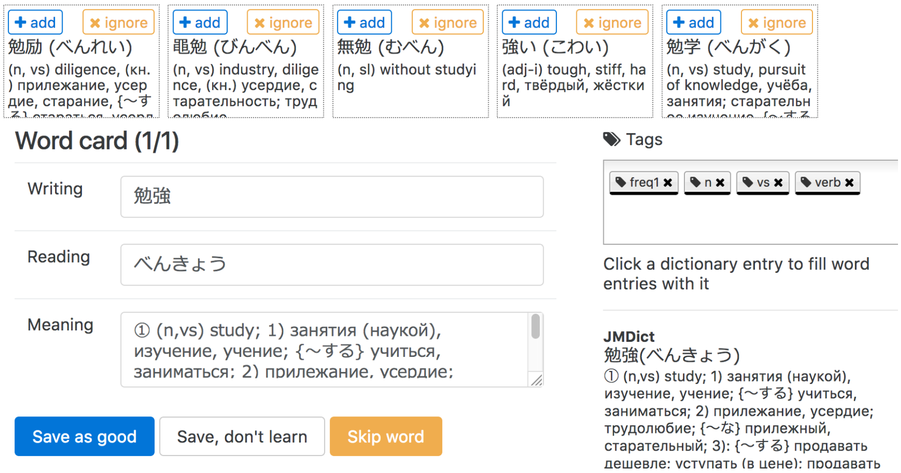
Figure 3 shows question and answer sides of a reading card for the word "盗む". Because it is a reading card, the question, which is an example sentence, has the target word written in hiragana. Hiragana reading is estimated using Kytea analyzer [4]. Answer cards for verbs has information about polite and -te form verb inflections which can be exceptions from usual rules. For example, 帰る having 帰ります polite form while 変える has 変えます.


To add new words enter them below, one per line.
耽る
漆黒
眠る

耽る	漆黒	眠る
ふける	しっこく	ねむる,ねぶる
to indulge in; to (	jet black	(v5r,vi) to sleep (

Figure 4: Fragment of word add form of Kotonoha

Process of adding new words to Kotonoha is designed to be batch by default. For example, in Anki users have to create cards one by one. In contrast, Kotonoha lets learners to input a list of words one per line. Figure 4 shows word add form of Kotonoha with three words to add. Screenshot shows that the learner has not added 耽る or 漆黒 before, but has already added 眠る. For the words, which are already added, Kotonoha give the learner an option to mark the word as forgotten (down arrow in the top-right of word card). In this case, the word will be scheduled earlier than spaced repetition algorithm had computed. Dictionary information for words is filled from JMDict [1]. It is a Japanese — multi-language dictionary. Examples on screenshots have glosses in English and Russian.



勉励 (べんれい)
(n, vs) diligence, (кн.) прилежание, усердие, старание, {〜する}

黽勉 (びんべん)
(n, vs) industry, diligence, (кн.) усердие, старательность; труд

無勉 (むべん)
(n, sl) without studying

強い (こわい)
(adj-i) tough, stiff, hard, твёрдый, жёсткий

勉学 (べんがく)
(n, vs) study, pursuit of knowledge, учёба, занятия; старательн

Word card (1/1)

Writing: 勉強

Reading: べんきょう

Meaning: ① (n,vs) study; 1) занятия (наукой), изучение, учение; {〜する} учиться, заниматься; 2) прилежание, усердие;

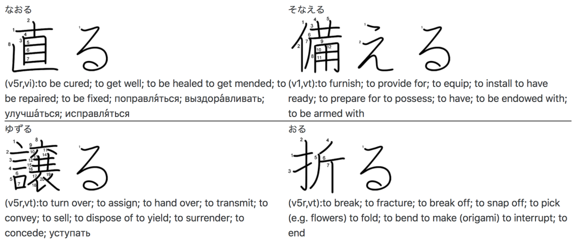
Tags

Click a dictionary entry to fill word entries with it

JMDict
勉強(べんきょう)
① (n,vs) study; 1) занятия (наукой), изучение, учение; {〜する} учиться, заниматься; 2) прилежание, усердие; трудолюбие; {〜な} прилежный, старательный; 3): {〜する} продавать дешевле: уступать (в цене): продавать

Figure 5: Word approval process

After adding words, a learner goes over the list once more, approving them for learning. Approval process is shown on the Figure 5. In this step, the learner can edit flashcard, assign tags or assign a different dictionary entry for the word if there are several ones with the same writing. Kotonoha also shows several recommended words containing same kanji characters which were not yet added by the learner, so those words could be easily added for learning as well.



なおる
直る
(v5r,vi):to be cured; to get well; to be healed to get mended; to be repaired; to be fixed; поправля́ться; выздора́вливать; улучша́ться; исправля́ться

そなえる
備える
(v1,vt):to furnish; to provide for; to equip; to install to have ready; to prepare for to possess; to have; to be endowed with; to be armed with

ゆずる
譲る
(v5r,vt):to turn over; to assign; to hand over; to transmit; to convey; to sell; to dispose of to yield; to surrender; to concede; уступать

おる
折る
(v5r,vt):to break; to fracture; to break off; to snap off; to pick (e.g. flowers) to fold; to bend to make (origami) to interrupt; to end

Figure 6: Writing practice output

In the learning process there could be words difficult to remember for the first time. Kotonoha supports outputting such words in a form suitable for writing practice with stroke order. Figure 6 shows several words in such form.

All these features makes Kotonoha a useful and interesting system for learning Japanese words.

4 Example Sentences

Example sentences for Kotonoha are added to cards completely automatically. They come from a web corpus which contains around 700M sentences. Extraction process is described in the thesis [7]. The process is done in two steps. In the first step, the system creates a search query based on the target word and its part of speech and makes search request to DepFinder [8] selecting up to 10k example sentence candidates. The second step selects sentences which are non-similar to each other, representative, and the target word should be the most difficult one in the sentence.

Example extraction was evaluated by Japanese learners. Each learner has reviewed example sentences by the system and two baselines for 14 words. For each word there was three lists of sentences presented: one for each method. Learners preferred the list produced by the system in 50% of cases.

Still, the evaluation was subjective and the process was very time consuming. Participants were selecting example sentences they like, not ones that would be helpful for remembering the target word.

In contrast to that, Kotonoha can be used to evaluate which example sentences were actually helpful for language learning. In spaced repetition participants report confidence of card knowledge for each card. We suppose that reported confidence would depend on multiple parameters of last repetition, including the shown example sentence. In the case when flashcards content was the same, with exception of example sentences, and students were of similar proficiency, we hypothesize that confidence scores could be used for evaluating whether the example sentence from *the previous* flashcard was helpful for remembering the word. Of course, to get the estimation statistically significant we need to have many learners using the system. Still, we believe that getting a large number of students to use a spaced repetition system is easier than performing a large scale experiment.

5 Limitations and Problems

Example sentences extraction uses JUMAN dictionary, which has different word segmentation compared to JMDict. Presently, example sentences are extracted for words which have one-to-one correspondence in JUMAN and JMDict.

For the test users of Kotonoha, example sentences could be found for approximately 70% of added words.

Words without examples could largely be divided into following groups.

- Words which consist of more than one JUMAN morpheme. This group has two subcategories.

  The first one contains mostly multi-morpheme expressions which are usually added into dictionaries with glosses, but are not useful in dictionaries of morphological analyzers. This group includes entries like "癇にさわる", "年末年始", "けじめを付ける" or "粗大ごみ". Handling multi-morpheme expressions of this kind is a future work for the example extraction.

  The second one are compound verbs which are divided into independent morphemes by JUMAN. The problem with compound verbs is that it is very difficult to balance semantic consistency and usefulness for search engines in one standard. Compound words which have finer granularity (are split into multiple morphemes) could be handled similarly to the first subcategory. Compound words which have rougher granularity require modification of either dictionary or tokenization process, which is not trivial.

- Words of unsupported parts of speech. In the present, example extraction supports only four parts of speech: nouns, verbs, adjectives and adverbs. Words like counter suffixes, while being very useful to learners, are not supported yet.

- Words which do not exist in JUMAN dictionary. Those words need to be added into the dictionary to be supported.

6 Conclusion

This paper presents a new spaced repetition system which uses automatically extracted example sentences to create question cards. In our best knowledge, this is the first spaced repetition application which automatically assigns example sentences from huge corpora to flashcards and shows a new example sentence for each repetition. Most of other apps use manually prepared example sentence corpora or show only a single example sentence for a word.

We theorize that it would be possible to use the system as a source of data for objective evaluation of example sentence quality. Also, the learning log from the system should be useful for general improvement of example sentence quality.

Lastly, example extraction and spaced repetition methods are language independent. The same approach could and should be done for other languages except Japanese.

References

[1] Jim Breen. JMdict: a Japanese-Multilingual Dictionary. In *COLING 2004 Multilingual Linguistic Resources*, pages 65–72, Geneva, Switzerland, August 2004.

[2] Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354–380, May 2006.

[3] Arthur W. Melton. The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 9(5):596–606, October 1970.

[4] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 2, Short papers*, pages 529–533. Association for Computational Linguistics, 2011.

[5] Philip I. Pavlik and John R. Anderson. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology. Applied*, 14(2):101–117, June 2008.

[6] Burr Settles and Brendan Meeder. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany, August 2016. Association for Computational Linguistics.

[7] Arseny Tolmachev. *Automatic Extraction of Diverse and High-quality Example Sentences from Large Scale Corpora for Language Learning*. Master's thesis, Kyoto University, Kyoto, 2016.

[8] Arseny Tolmachev, Morta Hajime, and Sadao Kurohashi. A grammar and dependency aware search system for japanese sentences. In *言語処理学会 第 22回年次大会 発表論文集*, pages 593–596, 仙台, mar 2016.

[9] Piotr A. Wozniak. *Optimization of learning*. Master's thesis, University of Technology in Poznan, Poznan, 1990.