Extraction of example sentences for improved reading and understanding of Japanese texts

Nahian Jahangir

June 11, 2015

Abstract

Properly learning and understanding foreign languages can be a difficult task. This is true for native English speakers who try to improve their Japanese reading comprehension. The difficulties of Japanese list as follows, from having three different writing systems to being a heavily-context based language. In this paper, I focus on solving the reading comprehension problem through the use of example sentences. Using a target word and its source sentence, I try to create a program that returns a proper example sentence that retains the context and contains easier grammar and terms to understand. There are three different approaches to this: simple overlap, collocations overlap, and weighting words. I evaluate each of the approaches based on their precision, recall, and mean average precision. The collocations approach does the best amongst the three, with its own advantages and disadvantages. I then discuss what the next steps of the project can be in creating proper example sentences

Contents

1 Introduction

Learning languages can be a difficult task for everyone just starting. An especially difficult language to learn for native-English speakers is Japanese, which the Defense Language Institute categorizes as Level IV in learning difficulty for native-English speakers. Language proficiency is evaluated in several different aspects, including reading, writing, speaking, and listening. Japanese reading comprehension in particular has several difficulties. There are three parts to the writing system hiragana, katakana, and kanji. Katakana and hiragana rely on their own different syllabic graphemes . Kanji are sets of characters that number in the thousands, each representing different words and terms, which can be combined to produce even more words and term. Furthermore, identifying words becomes much more difficult in Japanese writing as there are no spaces between words. Native speakers are able to identify words due to the fact that they are ingrained within the culture. They know the words of the language, they understand the sound and the meaning of the word, and thus, are able to read their native language quite well. [5] However, for novices learning the language, the challenge of reading comprehension becomes steep with all these obstacles: how do we overcome this?

When a non-native reader comes across an unknown word or phrase in some text, the logical step is to look up the word in a dictionary. However, one word can have multiple meanings depending on how it is used in the sentence. This is especially true in Japanese, as it is heavily based on the context of the sentence. An example of this is the Japanese word ”taihen.” Based in the context of the sentence, it has a variety of definitions. For example, the sentence ”kyou no shuukudai ga taihen desu” translates to ”Today’s homework is difficult,” where ”taihen” means diffiult. However, the sentence ”taihen benkyou ni narimashita” translates to ”You’ve taught me a lot,” where ”taihen” means a lot. For a novice in the language, it can be quite difficult to choose the correct meaning of the word without also having a firm understanding of the context at hand.

Using the dictionary to help the reader understand the word is a solution to a small problem, a solution that has its own obstacles. One such problem is that the word may have many definitions to the word. If the context is difficult to understand, finding the right definition becomes even more of a problem. Looking up the definition of the word also limits the learning of the language, short of memorizing the word and

its definition. There are still times where the reader may understand what the word means, but still not understand what the sentence is saying.

As a student studying the Japanese language during my time at college, I have come across my own issues when faced with unknown words. Though online tools and sources are a great alleviator in finding out the definitions of these words, it only helped answer questions rather than help supplement my Japanese knowledge, past memorizing the word and its definition. To understand the context and the entirety of word in the sentence, example sentences seemed like a step in the proper direction. However, very few online tools use example sentences and even fewer integrate them into their systems effectively. My vision for this program is so students can learn and understand these words in a way that emphasizes using more critical thinking.

Therefore, I propose a method for the reader that both improves their understanding of the text and strengthens their knowledge of the Japanese language. For unknown words and difficult sentences, it gives the reader easier example sentences based on the original sentence for better readability and understanding of the word and context at hand. Say for example, the reader is approached with this sentence: ”ryoushin ga nen wo tottara mendou wo miru tsumori desu”, which means ”In the case my parents get older with age, I will look after them.” If the reader is unable to understand what ’mendou’ means, they can look the word up in a dictionary and find several definitions for it; however, the word ’mendou’ has a completely different meaning when its directly affected by the verb ’miru.’ Using the program, the user would type in the word ’mendou’ and the sentence that it is in, ”ryoushin ga nen wo tottara mendou wo miru tsumori desu.” The program would then return another sentence that is both easier to understand and retains the same context, such as ”anata ga kaimono ni itteiru aida, kodomo no mendou wo mimashou,” which means ”While you go shopping, let us look after the children.” This example sentence uses basic grammar rules and has children as the subject, which helps drive the point that ’mendou’ in both sentences is used in the context of looking after people.

This method will be implemented through a computer program that takes in a target word and a source sentence. When the reader reads the text and comes across a word he or she does not understand, simply inputting it into the program should have an example sentence returned that helps them understand the

word within the simpler context.

2 Language and Related Works Overview

2.1 Japanese Language Ranked Category IV by the DLI

The Defense Language Institute (DLI) is an educational and research institution that provides linguistic and cultural instruction to the US Department of Defense. The mission of the institute is to "provide culturally based foreign language education, training, evaluation and sustainment to enhance the security of the nation" [1], in short, to help the American military train itself in languages around the world. The DLI categorizes all of the languages it teaches in four categories, where the category increments correspond to the difficulty a native English speaker would have learning the language. Category I languages are easier to learn quickly, and thus, take a shorter length of time, whereas Category IV languages are much more difficult, and thus, take a longer length of time to fully learn. The categories are listed here [1]:

- **Category I**: 26 week of courses, includes Spanish, French, Italian and Portuguese.
- **Category II**: 35 weeks of courses, includes German and Indonesian
- **Category III**: 48 weeks of courses , includes Dari, Persian Farsi, Russian, Uzbek, Hindi, Urdu, Hebrew, Thai, Serbian Croatian, Tagalog, Turkish, Sorani and Kurmanji
- **Category IV**: 64 weeks of courses, includes Arabic, Chinese Mandarin, Korean, Japanese and Pashto

As shown here, the Japanese language is characterized as Category IV in learning difficulty. There are many characteristics of the Japanese language that make it quite difficult for native English speakers to get a firm grip on the language, including its writing system and its ambiguous semantics.

2.2 Characteristics of the Japanese Reading and Writing Systems

The difficulty of learning languages depends on where the reader is from and where the language they are trying to learn is from. For native English speakers, languages that incorporate special characters and use a syllabic alphabet are the most challenging to learn, according to the Defense Language Institute.

As mentioned earlier, the Japanese language has three different systems in the overall writing system: hiragana, katakana, and kanji. Kanji are characters borrowed from China around the sixth century of any script of Japan's own. [4] Hiragana and katakana came later, the scripts had been developed in different parts of the country by around the ninth century, and this enabled for the first time the written representations of the pronunciation of Japanese words, inflections, and other manifestations of grammar. [4] Hiragana and katakana are called the kana systems, comprised of syllabic grapheme symbols. Hiragana is used to write native Japanese words and katakana is used to write words that are neither in Japanese nor Chinese. Kanji are characters that represent words and terms and number in the thousands; all kanji also have their own readings, known as furigana, basically the hiragana reading and writing of the kanji. The difficulty to read these characters is more pronounced as there is little to no spacing within the sentence. Having sentences made up of all three writing systems are useful in figuring out where the words start and end, but such cases aren't present all the time. Sometimes a sentence is entirely made up of hiragana, other times it is made up of several different, complex kanji word compounds. For these cases, there are not many clear cut clues to create inferences from the text, so knowing when a word begins and ends is exceedingly challengingly for a non-native speaker.

This problem is compounded when dealing with the semantics of the language. Semantics in language has many forms of ambiguity, particularly lexical ambiguity which deals with polysemous words. There are several examples in Japanese: "The verb miru appropriately illustrates the polysemous nature as well." [5] Its meanings include:

1. to look
    - Example: neko ga mimashita
    - Translation: I looked at the cat
2. to look after
    - Example: ryoshin ga nen wo tottara mendou wo miru tsumori desu
    - Translation: I'll look after my parents when they get old
3. to examine a patient

- Example: isha ga kanja wo miru
- Translation: A doctor examines a patient

Hence, native speakers of English will face many obstacles due to the difficulties Japanese reading comprehension presents to them.

2.3 Tanaka Corpus Background

I use the Tanaka Corpus as my database of example sentences to choose from. The Tanaka Corpus is a multilingual parallel corpus authored by Tanaka Yasuhito, a professor at Hyogo University in Japan. The beginnings of the project were based on the method of extracting Japanese-English bilingual newspaper articles and broadcast media news reports. Each of his student were charged with the task of acquiring 300 sentences pairs each, afterwards they would review the sentences collected by removing duplicates and correcting entries with errors. After 3-4 years, there were 212,000 sentence pairs within the parallel corpus with the following characteristics [2]:

- 40 percent of the sentences had a personal pronoun as the subject
- The average length of English sentences was 7.72 words, with the longest being 45 words
- Most often, sentences were everyday use sentences
- Interrogative and exclamatory sentences accounted for only 7.64 percent and 0.95 percent, respectively

There were several modifications done to the corpus as well:

- an initial regularization of the punctuation of the Japanese and English sentences was carried out, then duplicate pairs were removed, reducing the original file from 210,000 pairs to 180,000 pairs
- sentences which differed only by differences in orthography (e.g. kana/kanji usage, okurigana differences), numbers, proper names, minor grammatical points such as plain/polite verb usage, etc. were reduced to single representative examples

- sentences where the Japanese consisted of a short Japanese statement in kana were removed
- sentences with spelling errors, kana-kanji conversion errors, etc. were corrected
- sentences where the Japanese was too garbled to derive a valid English equivalent were removed.

The Tanaka Corpus was one of the few corpora in its time to ever compile a comprehensive database of sentences .Jim Breen realized the potential of the corpus as a source of example sentences for his electronic dictionary server. Each dictionary entry consists of kanji elements, reading elements, general coded elements, and the sense elements. He edited, reformatted and indexed the corpus and linked it at the word level to the dictionary function in the server. In the same vein, I plan to use this corpus as a source for example sentences with my program. Rather than incorporate it into a dictionary server, I plan to use the sentences directly. [3] Currently, the corpus is being maintained by the Tatoeba project, which has been extended further to incorporate more sentences of many different languages. I will be using the Japanese sentences as a source of example sentences within the corpus.

3 Example Sentence Extraction Approaches

Japanese reading comprehension is marred by difficulties in identifying and understanding words in the text. Though looking the word up in the dictionary is one solution, it can often take time and serve little to no better understanding to the context at hand. Therefore, easier example sentences of the word using the same context should serve as a better alternative to understand and learn Japanese.

Good example sentences should accurately reflect the sense of the word in the original sentence as well as being easier to read for novices. The baseline for the approach is finding a sentence in the corpus that has the the highest number of overlapping words with the original sentence. The next approach is the collocations of sentences method, which is to find a sentence in the corpus a sequence of words or terms that co-occur more than often throughout the corpus. The third and last approach is the weighted word method, which is to give each word a weighted value based on how often they're found throughout the corpus; it finds the sentence whose accumulative weight is the highest.

3.0.1 Simple Overlap Comparison Algorithm

Algorithm 1 Find the sentence with the most overlap

```
1: bestscore = 0
2: example sentence = ""
3: source vector = vectorCreation(sentence)
4: for other sentence in corpus do
5:    other vector = vectorCreation(other sentence)
6:    if word in other vector then
7:       score = compareOverlap(source vector, other vector)
8:    end if
9:    if  score > best score then
10:      best score = score
11:      example sentence = other sentence
12:   end if
13: end for
14: return  example sentence
```

The baseline of this program simply takes in a word and a sentence and returns an example sentence. It creates sets out of the source sentence and for each of the sentences within the corpus. It then creates a score of how much overlap the source sentence has with the example sentence. The overlap score is the size of the intersection between the two sets of the source sentence and potential example sentence. The higher the score, the more of an overlap there is. The example sentence with the best score is returned and used as an example sentence.

This is a good starting point in trying to find the best example sentence; however, there are several issues with this. For starters, longer sentences are more likely to get a high score than that of smaller sentences since they give more opportunities for words to overlap with each other. Here are some examples:

"kyou, subete wo ohanashi suru wakeniha ikimasenga, yousuruni, watashi no mendou ha, kotoshi no natsu, touchini ha konai no desu."

This sentence was ranked 6th: not only was it missing the context of the original sentence, it contained many different complicated words and grammar. Though it was not a proper sentence, the length of the sentence gave it the boost in score to attain a better placement within the rankings.

Since long sentenes have unfair advantages, the scores of all the sentences were normalized. These normalized scores took into account the length of the original sentence and made it less of a significant

factor in calculating the final score.

Counting the overlap is the first and foremost aspect of the algorithm to change since the best example sentence the program will ever return is one that is exactly like it. As mentioned before, a good example sentence should reflect the original sense and give the reader an easier time to understand the word and the context it is in. Therefore, I go into a couple of methods which steer past the baseline, the first of which is the ”Collocations Overlap Comparison” method.

3.0.2 Collocations Overlap Comparison Algorithm

Algorithm 2 Find the sentence with the most overlap in terms of collocations

```
extracted sentences = []
ranking = []
source vector = vectorCreation(sentence)
dictionary =
for other sentence in corpus do
    other vector = vectorCreation(other sentence)
    if word in other vector then
        extracted sentences.add(other sentence, other vector)
        for words in other vector do
            if word in dictionary then
                dictionary[word] = dictionary[word]+1
            else
                dictionary[word] = 1
            end if
        end for
    end if
end for
dict set = convertDictionaryToSet(dictionary)
removeStopCharacters(dict set)
sort(dict set)[:20]
for sentence and set in extracted sentences do
    score = compareOverlap(dict set, other vector)
    ranking.add(sentence, score)
end for
sort(ranking)
return ranking
```

This update to the program returns a ranking of quality example sentences rather than just the best. This was implemented in order to gauge how well the system was doing as a whole with all the example

sentences at its disposal.

The collocations method does not use the source sentence for its calculations. Rather, it uses the words that surround the target word for its calculations. The program does this by creating a count for each of the words in each of the sentences within the corpus. The sentences it chooses are the sentences that contain the target word. After it accumulates the entire list, the 20 most popular words are taken and placed in a set.The comparison overlap calculation is then applied upon each sentence within the corpus (that contain the target word) with the ”popular set”. A ranking of the sentences is then found based on the overlap scores.

Stop characters were introduced to this algorithm as well. These characters are removed from the sentences based on their significance in the meaning of the sentence. In Japanese, particles are one or more Hiragana characters that attach to the end of a word to define the grammatical function of that word in the sentence. Here is an example of particles, which are highlighted **bold** below:

”watashi **ga** gaisshutsu shiteiru **aida**, inu **no** mendou **wo** mitekurenai”

Particles occur in all sentences, but their significance to the context of the sentence isn't needed. Therefore, I decided to try removing particles from sentences in order to get a more precise overlap score between sentences.

The second method I go into differs from simply counting the overlap. This approach is called the ”Weighted Weights” method, which approaches the words and sentences in a slightly different manner.

3.0.3 Weighted Words Algorithm

This approach tackles the problem by assigning a weight to all of the words and finding the example sentence that has the highest weight amongst all of them. The weight is assigned based the number of time a word appears divided by all the sentences that include the target word from the corpus. This is implemented by having a separate weighted dictionary that has all the terms and their relative scores. In calculating the score, weighted dictionary is referred to when finding the score of a word and it is then summed up to the sentence's total score. This method as well takes into account the normalization of sentences and removing stop characters from the sentences.

Algorithm 3 Find the sentence with the highest weight

```
1: extracted sentences = []
2: ranking = []
3: source vector = vectorCreation(sentence)
4: dictionary =
5: for other sentence in corpus do
6:     other vector = vectorCreation(other sentence)
7:     if word in other vector then
8:         extracted sentences.add(other sentence, other vector)
9:         for words in other vector do
10:            if word in dictionary then
11:                dictionary[word] = dictionary[word]+1
12:            else
13:                dictionary[word] = 1
14:            end if
15:        end for
16:    end if
17: end for
18: for term, count in dictionary do
19:    weigthed score = calculateScore(term, count)
20:    weighted dictionary[term] = weighted score
21: end for
22: for sentence and set in extracted sentences do
23:    removeStopCharacters(other vector)
24:    score = compareOverlap(weighted dictionary, other vector)
25:    ranking.add(sentence, score)
26: end for
27: sort(ranking)
28: return ranking
```

4 Information Retrieval Evaluations

The evaluations of each of these methods is based on the same evaluations done upon information retrieval devices. The precision, recall, and mean average precision was calculated for each of these methods. The precision and recall were calculated by the following formulas:

precision = number of selected sentences that are good examples / number of all selected sentences

recall = number of selected sentences that are good examples/ number of all good examples

The number of all selected sentences was the top 20 example sentences chosen by the algorithm. Good examples were example sentences chosen out by me, based on a beginner's knowledge of Japanese.

4.1 Baseline

Words	Precision	Recall	Average Precision
″mendou″	2/20	2/9	.0687
″tanomu″	2/20	2/3	.0582
″yasui″	2/20	2/11	.0282

Mean Average Precision- .0517

4.2 Collocations Method

Words	Precision	Recall	Average Precision
″mendou″	2/20	2/9	.2275
″tanomu″	2/20	2/3	.0349
″yasui″	2/20	2/11	.0714

Mean Average Precision- .1112

4.3 Weighted Words Method

Words	Precision	Recall	Average Precision
″mendou″	3/20	3/9	.1221
″tanomu″	1/20	1/3	.0109
″yasui″	1/20	1/11	.0109

Mean Average Precision- .0479

5 Conclusions

The method with the highest mean average precision was the Collocations method, at around .1112, compared to that of the Baseline's method, at around .0517, and the Weighted Words Method, at around .0479.

The Collocations method was both precise and intuitively makes sense in developing good example sentences. The system scores sentences that contain commonly written phrases higher, and since the corpus is made up of everyday Japanese sentences, you would get more of those types of sentences. Having understandable, commonly used sentences as example sentences can provide the user further understanding and learning of the Japanese language.

Though the Baseline method dealt with using the source sentence, both the Collocations and Weighted Words method dealt moreso with the contents of the corpora. The Collocations method gave higher scores based on how common a group of words were throughout the corpus. The Weighted Words method gave words higher weight based on how much they occurred throughout the corpus as well. These methods are heavily based on corpora and can produce biased results.

For example, say the sentence we are reading is "Montgomery fished from the **bank** from dusk till dawn, promising he wouldn't leave until he caught the fabled fish." If we don't understand is the word "bank," we could hopefully use the program to get an example sentence that helps us understand it. However, if the corpus the program uses only understands the word "bank" as an institution to withdraw and deposit money from, then whatever example sentence we get from the program will ultimately be useless.

Improvements that can be made to the system can be an incorporation of a kanji proficiency system. This would limit the number of complicated, lengthy sentences in our system, leaving it with user appropriated sentences. Another is expanding the database of example sentences for the program. Having a variety of sentences to choose from can help the program pinpoint more and better example sentences to choose.

References

[1] The defense language institute. http://new.dliflc.edu/. Accessed: 10/05/2015.

[2] Electronic dictionary research and development group. http://www.edrdg.org/wiki/index.php. Accessed: 21/11/2014.

[3] Jim breen's edict web page. www.csse.monash.edu.au/ jwb/wwwjdic.html. Accessed: 20/11/2014.

[4] Nanette Gottlieb. *Language and Society in Japan*. Cambridge University Press, February 2005.

[5] Natsuko Tsujimura. *An Introduction to Japanese Linguistics*. Wiley-Blackwell, September 2006.