Automatic Generation for Mandarin Vocabulary Practices

Fabian Lang
Department of Media Science
University of Tübingen
f.lang.1993@web.de

Chih-Chun Chang
Department of Computational Linguistics
University of Tübingen
jerrychihchun@gmail.com

Abstract

For our generation task we designed a novel system that can be used for Mandarin vocabulary and text comprehension practice. Questions in the form of cloze sentences can be generated from any texts. The study includes evaluations on questions generated by the system for texts of three levels of proficiency and discussions on the results, framework limitations and future research. The results show that texts of the lowest level A have the highest rate of question generation, while those of level C have the best quality of questions. The system is intended to be used by teachers for questions regarding vocabulary knowledge and practices.

1 Introduction

We build our system to create automatic generated 2-choice questions for Mandarin vocabulary practice inspired by Susanti et al. (2015) on English vocabulary tests . The option pairs consists of the blanked word and an antonym of itself. The idea is to provide a lexically informed framework that can be used for teaching and learning vocabulary.

We constructed antonymous distractors based on the information derived from synsets contained in the Chinese WordNet (CWN[1]) (Huang and Hsieh, 2010). The CWN synsets contain information about hypernyms, part-of relations, synonyms, and antonyms.

The pre-processing of our input is done with NLPIR (Natural Language Processing Information Retrieval), which includes word segmentation, POS-tagging and unknown word detection.

The input and output data are illustrated before the evaluation and results. We also discuss our observations, limitations, and future work before summary. Materials and source codes are provided at the end of the paper.

2 Automatic Generation

2.1 Text Pre-processing

Before processing the data with our system a crucial initial step is to transform all Chinese characters to one unified form. The reason for this is to allow the system to universally process any text resource, and not to restrict itself to texts containing characters from either Traditional Chinese (e.g. in Taiwan and Hong Kong) or Simplified Chinese (e.g in Mainland China). Furthermore, this step is necessary in order to avoid incompatibility between our systems components, since the information contained in the CWN is solely encoded in traditional characters, the segmentation software however just processes simplified characters.

For the sake of simplicity we transform any input to Simplified Chinese using a Python module Hanziconv[2]. Once the input is available in its simplified form we move on and segment the input using a Python module called PyNLPIR[3] which runs on the popular NLPIR/ICTCLAS segmentation software.

2.2 Cloze Generation

PyNLPIR also allows for detection of keywords, which we use to generate suitable clozes. For every text we extract all available keyword information and use it to match corresponding CWN synsets. As already mentioned, CWN synsets contain information about synonyms, antonyms, hy-

[1] Online queries can be accessed at http://lope.linguistics.ntu.edu.tw/cwnvis/

[2] https://github.com/berniey/hanziconv

[3] https://github.com/tsroten/pynlpir

ponyms, and hypernyms. Once the system acquired all possible antonym synset pairs of a given keyword in a sentence, it transforms the sentence using RegEx and creates a blank consisting of multiple underscore characters and two valid options A and B below the question.

The A and B options are randomized so the learner doesn't know which one is the distractor and which one is the correct answer for each of the questions. When selecting a distractor choice the system favours same character length antonyms over different character length ones. This is a simple method to display more accurate antonym distractors. The antonym choice also undergoes a randomization process and once a keyword has been successfully replaced in a sentence, subsequent sentences replacing the exact same word only prompt the remaining antonyms, but not previously used ones.

2.3 Algorithm

The pseudocode for our generation is tabulated into steps in Table 1. More information on the set up is provided in the last section.

```
pre-process(t)
    simpl_char = hanziconv(t)
    segm = pynlpir(simpl_char)
    sentence = split_strip(segm)
    return sentence
generate-cloze(t)
    kws = pynlpir_keywords(t)
    retrieve_cwn_antonyms(kws)
    sentence = pre-process(t)
    for word in sentence:
    if antonym in sentence:
    regex(word,’____’,sentence)
    ant = same length first
        otherwise
    ant = different length
    shuffle word,antonyms
    remove used antonym
    endif
    endfor
    return cloze-sentence
```

Table 1: Pseudocode of cloze generation.

3 Data

3.1 Input

The input for our vocabulary practice generation is derived from the online mock exams of Test of Chinese as a Foreign Language (華語文能力測驗 Huáyǔwén Nénglì Cèyàn, TOCFL) [4]. The test is available in three bands, A, B, and C. Depending on the scores, test takers are further divided into two levels for a total of 6 levels from A1 through C2.

Our texts are arranged according to the three bands A, B, and C, containing 15, 28, 19 passages respectively. Their average word count in each passage is 140, 242, and 382.

3.2 Output

The three punctuation symbols period, exclamation mark, and question mark are used as sentence delimiters (。,！,？). For each band, 51, 125, and 63 questions are generated. Each question comes with a pair of options with the blanked word and an antonym. A snippet of an example generated text is shown in Figure 1 with the text and question translated below. Lexical information regarding synonyms, antonyms, hyponyms and hypernyms are printed at the user's disposal for additional information. More details will follow in the next section.

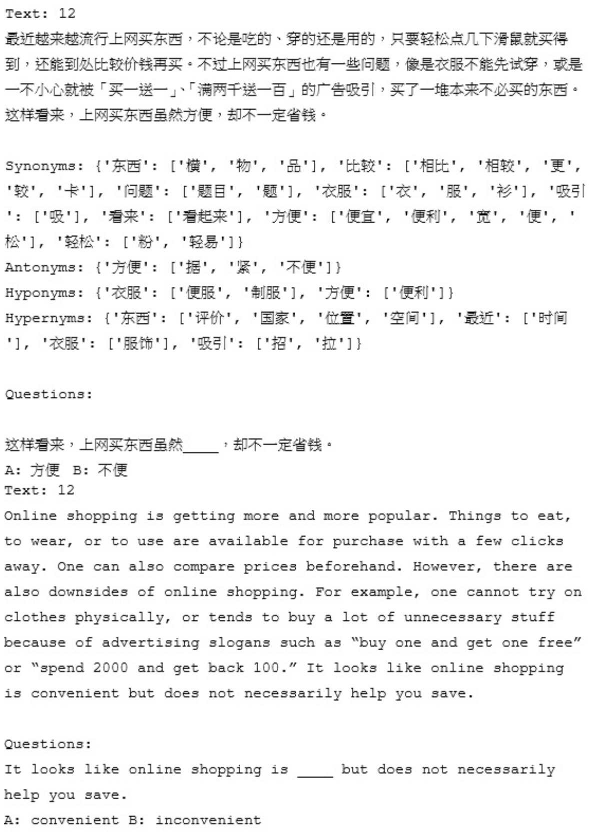
4 Evaluation & Results

We generated 2-choice questions for texts from the three levels/ bands. Before the cloze practices, lexical information such as synonyms, antonyms, hyponyms, and hypernyms are printed at the user's disposal. In case the system fails to generate a suitable question, this provides instructors lexical information for manual generation.

4.1 Criteria

Questions generated in the study were rated by the second author as a native speaker of Mandarin. Each was rated as being good, neutral, or bad with detailed descriptions in Table 2. Scores as 1, 0, and -1 respectively, were assigned and tallied. A ternary rating system is to determine good questions with syntactically compatible and semantically contrastive pairs, neutral questions if meeting only one requirement, and bad questions if the

[4] For more information, please visit https://www.sc-top.org.tw/index.php



```
Text: 12
最近越来越流行上网买东西，不论是吃的、穿的还是用的，只要轻松点几下滑鼠就买得
到，还能到处比较价钱再买。不过上网买东西也有一些问题，像是衣服不能先试穿，或是
一不小心就被「买一送一」、「满两千送一百」的广告吸引，买了一堆本来不必买的东西。
这样看来，上网买东西虽然方便，却不一定省钱。

Synonyms: {'东西': ['横', '物', '品'], '比较': ['相比', '相较', '更',
'较', '卡'], '问题': ['题目', '题'], '衣服': ['衣', '服', '衫'], '吸引
': ['吸'], '看来': ['看起来'], '方便': ['便宜', '便利', '宽', '便', '
松'], '轻松': ['粉', '轻易']}
Antonyms: {'方便': ['据', '紧', '不便']}
Hyponyms: {'衣服': ['便服', '制服'], '方便': ['便利']}
Hypernyms: {'东西': ['评价', '国家', '位置', '空间'], '最近': ['时间
'], '衣服': ['服饰'], '吸引': ['招', '拉']}

Questions:

这样看来，上网买东西虽然____，却不一定省钱。
A: 方便  B: 不便
Text: 12
Online shopping is getting more and more popular. Things to eat,
to wear, or to use are available for purchase with a few clicks
away. One can also compare prices beforehand. However, there are
also downsides of online shopping. For example, one cannot try on
clothes physically, or tends to buy a lot of unnecessary stuff
because of advertising slogans such as “buy one and get one free”
or “spend 2000 and get back 100.” It looks like online shopping
is convenient but does not necessarily help you save.

Questions:
It looks like online shopping is ____ but does not necessarily
help you save.
A: convenient B: inconvenient
```

Figure 1: An example generated text.

pair is generated due to false lexical information or simply doesn't fit in the cloze.

4.2 Results

Results are presented and briefly discussed for each band in this section. A statistical overview is shown in 2.

4.2.1 Band A

The size of text for band A is 2101 with an average word count of 140 per passage. 51 questions were generated, marking the highest generation rate among the three bands. However, the averaged score at 0.138 (1 being good, -1 being bad) is the lowest among the three.

4.2.2 Band B

The size of text for band B is 6771 with an average word count of 242 per passage. A total of 125 questions were generated. The average question score is 0.528 (1 being good, -1 being bad), slightly lower than that of band C.

4.2.3 Band C

The size of text for band C is 7252, slightly larger than for band B with an average word count of 382 per passage. 63 questions were generated with the highest score average 0.540 among the three (1 being good, -1 being bad).

Type	Score	Descriptions
Good	1	both options share the same syntactical category and are semantically contrastive
Neutral	0	the antonym is of a different syntactic category or is not an antonym distinguished from its counterpart; questions are context-dependent
Bad	-1	both options fit or don't fit in the cloze; questions are generated due to incorrect parsing or tagging; lexical information is faulty

Table 2: Question evaluation.

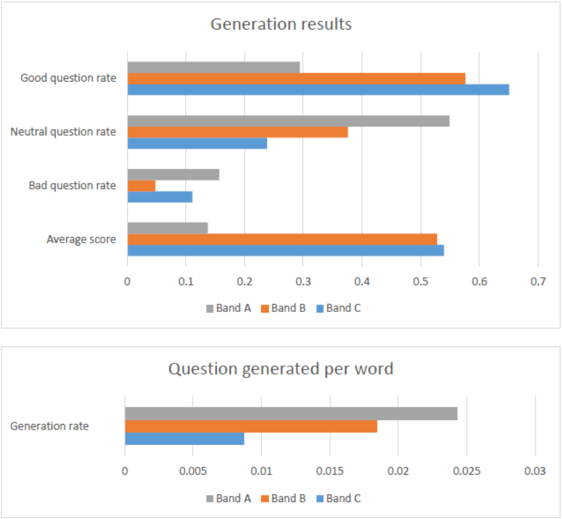


Figure 2: Results of generation.

5 Discussions

5.1 Observations of Generation

At a glance, an inverse relationship is revealed from the results between the generation rate and the average score.

On the one hand, band A has the highest generation rate in terms of the number of words. This could be due to its simplicity in its structure and rich lexical information as the words that appear are suited for beginners. On the other hand, band

A has the worst performance in terms of question scores. The chance of generating a good questions is only about 30%.

The opposite is the case for band C, which has the lowest generation rate but the highest average score. Its structural difficulty or use of advanced vocabulary might be the cause of the paucity of derived lexical information.

5.2 Framework Limitations

Heaton (1975) provides guidelines for creating English vocabulary tests, including: (1) each option should belong to the same word class, (2) the answer and distractors should be at about the same level of difficulty, and (3) all options should be of about the same length. In the light of (1), being a good question requires both items to be syntactically compatible for the category. If the antonym option does not fit in the cloze grammatically, it needs to be demoted to a neutral question. For (3) and also for stylistic and metrical matters, the system prioritizes antonyms sharing the same length of characters as the item being blanked. This also helps us to eliminate options whose antonymous relation to the item is fairly weak or obsolete. However, (2) was not taken into account in our work. Incorporating frequency queries into the CWN would help us select level-appropriate items.

Another limitation of our system is related to parsing. A setback was posed on band A due to the failure to recognize and parse names correctly. One particular passage with 7 generated questions scored -4 with 4 bad questions. Had *Chén xiānshēng* 'Mr. Chen' been correctly tokenized and parsed, our system wouldn't have unsolicitedly generated adjectival antonyms for chén which has an archaic meaning for *old*. One surely wouldn't want to have a cloze on the surname of someone and further adjectivize it during vocabulary teaching and learning.

A third major limitation we have observed has to do with the structure of lexical information. Originally we experimented with two ways of cloze generation. The other unpresented system replaced the blanked word with a synonym when the word of the same keyword was targeted in subsequent clozes. The motivation was to provide help with the acquisition of its synonyms for the learner. However, the CWN also contains some faulty lexical information. For example, *chénggōng* 'success' unfavorably contains synonyms such as *National Cheng Gong University*, leaving our generation coarse. Therefore, it was discarded. This faulty quality also led to some bad questions, for example, *niánlíng* 'age' containing *fake* as a synonym, further giving *true* as an antonym in the generation. Most of the time it was such nonsensical pairs that gave rise to bad questions. Another issue that surfaces only occasionally is with polysemy. Under such circumstances, the antonym item might be matched with a sense that contextually is not strongly antonymous or does not share the syntactic category, leading the question to be not good.

5.3 Future Work

Considering our "one small step" for automatic generation of Mandarin vocabulary practices, there are many possible forms of improvement and extensions that can be added to this work in the future. As we have seen some limitations of our work in the previous subsection, automatic generation of vocabulary practices of Mandarin is actually quite promising.

With more robust lexical information and the improved tokenizers and parsers for Mandarin, vocabulary items should be able to be generated with respect to the three variables of item analysis formulated by (Gronlund, 1982): difficulty, discriminating power, and distractor effectiveness. Including items used in relevant contexts and at appropriate levels of difficulty, the generation system of Mandarin vocabulary is to grow towards becoming full-fledged.

6 Summary

We extracted texts from TOCFL reading comprehension templates for our study. We generated 2-choice vocabulary practices by text-preprocessing done with NLPIR and creating vocabulary options based on the lexical information retrieved from Chinese WordNet. Texts of level A have the highest generation rate, whereas texts of level C have the best average question score. We discussed these observations, limitations, and future research directions. With the constantly developing parsers, a more solid WordNet could effectively enhance the performance of generation with more distractor items of relevant contexts and appropriate difficulty.

References

Norman Edward Gronlund. 1982. *Constructing achievement tests*. Prentice Hall.

John Brian Heaton. 1975. *Writing English language tests: a practical guide for teachers of English as a second or foreign language*. Longman Publishing Group.

Chu-Ren Huang and Shu-Kai Hsieh. 2010. Infrastructure for cross-lingual knowledge representation-towards multilingualism in linguistic studies. *Taiwan NSC-granted Research Project (NSC 96-2411-H-003-061-MY3)* .

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic generation of english vocabulary tests. In *CSEDU (1)*. pages 77–87.

Supplemental Material

We attach the results of band A and our code file named run.py.

In order to get the system to work there are a few packages that need to be installed beforehand. These steps are as follows.

1) Place the run.py and cwn_dirty.csv in the same folder together. The cwn.py file needs to be moved to Python3's site-packages folder.[5]

2) CMDs to install Python packages
pip install hanziconv
pip install pynlpir
pynlpir update

3) In the same folder of run.py and cwn_dirty.csv there needs to be a text file containing the Chinese text resources. The text file requires to only have a single line. The line allows for multiple texts separated by | as the delimiter.

4) Usage: python run.py text.txt
The output text file result_text.txt will be written into the same directory.

[5] The cwn_dirty.csv and cwn.py can be downloaded from https://github.com/loperntu/cwn2/tree/master/cwn2