

Data Coverage vs. Data Size: A comparison of two large-scale corpora in Collocation Suggestion for Japanese Second Language Learners

LIS W. K. PEREIRA ERLYN MANGUILIMOTAN YUJI MATSUMOTO

Nara Institute of Science and Technology

{lis-k,erlyn-m,matsu}@is.naist.jp

1 Introduction

Natural language research ranging from automatic part-of-speech (POS) tagging to statistical machine translation (SMT), and the development of NLP applications such as grammatical error correction, utilize large amount of text. The assumption is that, a large-scale text can provide broader knowledge of the language that is useful in designing language models. Results from various researches show that by increasing the size of the data the performance scores also increase. Increasing the size of the training data in a spelling correction research improved the accuracy [1]. [2] mentioned that the size of a corpus is important in a phrased-based SMT. However, the improvement in the system is more evident for correcting errors such as articles, prepositions, and adjective errors, while not much improvement is seen for correcting agreement and tenses. Very large corpora are still necessary to obtain extensive information of the language, i.e. grammar; however is it enough to extract more complex information, such as the case of collocations? In this paper, we would like to show that data coverage of the corpora, more than corpora size, contributes to a better performance score in automatic collocation suggestion for Japanese second language learners. For collocation error suggestion/correction, enough coverage is relevant since we are dealing with substitution of open class words (nouns, verbs, adjectives and adverbs). We run two large-scale corpora, a) Mainichi Shimbun [3], and b) Balanced Corpus of Contemporary Written Japanese (BCCWJ) [4] in computing the collocation scores. We try to show that the data coverage of BCCWJ gives better collocation suggestion compared to the larger sized Mainichi Shimbun.

2 Collocation Suggestion Task

Collocations can be generally defined as conventional word combinations in a language. In Japanese language, some examples of collocations are: “ ocha wo

ireru (お茶を入れる)”, “ kusuri wo nomu (薬を飲む)” and “ yume wo miru (夢を見る)”. Learning collocations can be a very challenging task for second language learners since there are no well-defined grammar rules to determine collocation preferences. For instance, a Japanese language learner may write: 一昨日の夜、クラスメートの夢をした。(Ototoi no yoru, kurasumeeto no yume wo shita.) However, the correct sentence should be: 一昨日の夜、クラスメートの夢を見た。(Ototoi no yoru, kurasumeeto no yume wo mita .) Although both sentences are syntactically correct, the first one sounds unnatural, “ yume wo suru (夢をする)”. For our collocation suggestion system, the task is to suggest “ noun wo verb (noun-を-verb)” collocations to Japanese second learners. Given a learner’s construction, a set of word candidates (e.g. verbs) is suggested using word similarity algorithms and collocation measures. In computing word similarity, a confusion set derived from a Japanese language learner corpus (Lang-8)¹ is used [5]. This confusion set is for generating the verb suggestion candidates. Using two large-scale Japanese corpora (Mainichi Shimbun and BCCWJ), we compute the collocation scores using Weighted Dice coefficient [6]. Based on the confusion set and the collocation scores computed, the best candidates are ranked and suggested to the learner.

3 Corpora for Collocation Extraction

To show the difference in computing collocation scores using a wider coverage data and larger-sized data, we used a Japanese Newspaper Corpus, Mainichi Shimbun and BCCWJ, a balanced corpus that consists of books, magazines, newspapers, and many others [4]. The difference between these corpora is that, Mainichi Shimbun has a vocabulary coverage that a Japanese language learner may not have acquired. On the other hand, BCCWJ’s varied type

¹Lang-8: <http://www.lang-8.com>

Data	Mainichi Shimbun		BCCWJ
Size	1 year data 1991	2 year data 1991-1992	56,561 samples
noun を verb pairs	224,185	396,313	194,036
unique nouns	37,300	56,222	43,243
unique verbs	16,781	24,773	18,212

Table 1: Data Size Specification

Precision	Recall	F-Score
$\frac{tp}{tp+fp}$	$\frac{tp}{tp+fn}$	$\frac{2*precision*recall}{precision+recall}$

Table 2: Evaluation Metrics

of data source results to a wider vocabulary range. Table 1 shows the data size used in computing collocation scores and the words extracted, together with the "noun wo verb (noun を verb)" pairs and the number of unique nouns and unique verbs.

4 Experiment

First, we computed collocation scores using one year (1991) Mainichi Shimbun data. We then increased the data by another year (1992) of Mainichi Shimbun data as shown in Table 1. Lastly, we computed collocation scores using the BCCWJ data, choosing portions of the corpus that include: magazine, newspaper, textbooks, and blog data. For defining the collocation candidate set, we selected all the "noun wo verb (noun を verb)" tuples that co-occur at least three times in the corpus ($f \geq 3$). In theory, any pair of words that co-occur at least twice in a corpus is a potential collocation.

4.1 Evaluation

For the evaluation, a test data was created by crawling the revision log of a language learning SNS, Lang-8. It contains tuples of learner's sentence and its correction given by native speakers of Japanese language. The test set was constructed extracting the "noun wo verb (noun を verb)" tuples with incorrect verbs and their correction given by native speakers of Japanese language. In total, 269 tuples were selected and evaluated. We compared the verbs suggested by the system with the human suggested verb in the Lang-8 data. A match would be counted as a true positive (tp). A false positive (fp) occurs when the system can offer suggestions, but none of them matches the human suggested verb in the Lang-8 data. A false negative (fn) occurs when the system cannot offer any suggestion. The metrics we used for the evaluation are: precision, recall and F-score. The formula are shown in Table 2.

4.2 Experiment Results

Table 3 shows the recall, precision and f-score obtained using the different corpora. K-best evaluation was used to check the precision rank of the correction given by the system. For instance, *k-best 1* means that the correction given by our system was suggested in first place, *k-best 2* means that the correction was ranked either as first or second place. *K-best 5* means that the correction was ranked within the first five suggestions, and so on. The highest values are shown in bold type. Regarding the difference between the two sizes of Mainichi Shimbun data used, the smaller data (1 year data) obtained better precision rate than the bigger data (2 year data). However, when we use 2 year data of Mainichi Shimbun, there is an improvement in the recall rate, but the precision rate decreased. However, using BCCWJ, we could obtain the highest values for all the metrics: precision, recall and F-score.

5 Discussion

The main reason that even increasing the size of Mainichi Shimbun data used we could not obtain the same results when using BCCWJ is that there is a large gap between the vocabulary used in newspaper and the learner's vocabulary. For instance, when using Mainichi Shimbun, the recall rate is lower than when using BCCWJ because the correct collocation suggested in the learner corpus could not be found when computing collocation scores. However, since BCCWJ is a balanced corpus, covering a wide range of vocabulary, it was easier to find the correct collocation suggested in the learner corpus when computing collocation scores. Regarding the precision rate, BCCWJ gave higher ranks to the collocation correction, obtaining a higher precision rate compared to Mainichi Shimbun. That occurred because for many cases, Mainichi Shimbun assigns higher collocation scores to expressions that are not common in second learners' writing, and assigns lower scores to common expressions in their writing.

6 Conclusion

In this paper, we showed that data coverage of the corpora, more than corpora size, contributes to a better performance score when applied in automatic collocation suggestion for Japanese second language learners. In our experiments, we found that even increasing the size of newspaper data used could not outperform the results obtained with the use of a balanced corpus BCCWJ.

Mainichi Shimbun		BCCWJ		Mainichi Shimbun		BCCWJ	Mainichi Shimbun		BCCWJ
1 year (1991)	2 year (1991-1992)	56,561 samples		1 year (1991)	2 year (1991-1992)	56,561 samples	1 year (1991)	2 year (1991-1992)	56,561 samples
Recall			K-best	Precision			F-Score		
0.8513	0.9070	0.9702	1	0.6506	0.5819	0.6973	0.7375	0.7090	0.8114
			2	0.8165	0.7213	0.8544	0.8335	0.8035	0.9086
			3	0.8864	0.8360	0.9118	0.8685	0.8701	0.9401
			5	0.9519	0.9303	0.9770	0.8988	0.9185	0.9736
			10	0.9956	0.9918	0.9961	0.9178	0.9475	0.9830

Table 3: Recall, Precision and F-Score of Collocation Suggestion using different corpora

References

- [1] Michele Banko and Eric Brill. 2001, *Scaling to Very Very Large Corpora for Natural Language Disambiguation*, Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 26-33, 2001.
- [2] Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata (NTT) and Yuji Matsumoto, *The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings*, In Proceedings of the 24th International Conference on Computational Linguistics, pages 863-872 2012.
- [3] Mainichi Newspaper Co., Mainichi Shimbun CD-ROM, 1991-1992.
- [4] Kikuo Maekawa, *Balanced corpus of contemporary written japanese*, In Proceedings of the 6th Workshop on Asian Language Resources (ALR), pages 101–102 2008.
- [5] Lis Weiji Kanashiro Pereira, Erlyn Manguilimotan, and Yuji Matsumoto, *Collocation Suggestion for Japanese Second Language Learners*, 情報処理学会研究報告第 210 回自然言語処理研究会, Vol.2013-NL-210, 2013.
- [6] M. Kitamura and Y. Matsumoto, *Automatic extraction of translation patterns in parallel corpora*, In IPSJ, Vol. 38(4), pp.108-117, April 1997, 1997.