Iterated collocation extraction through mutual information in Mandarin Legal Documents

Anonymous ACL submission

Abstract

Collocation identification is an important dimension for multiple natural language processing tasks. In Mandarin, due to the orthography and the highly compositional nature, identifying collocations is especially challenging. While most popular segmentation tools can identify common collocations, their performances are largely sabotaged when dealing with domain-specific texts. In this paper, we present a novel collocation extraction technique aimed at domain-specific texts through iterated segmentation based on the popular mutual information measure and its other variant, averaged mutual information. It has been found that while mutual-information-based collocation extractions did not benefit from iterated segmentation, collocation extractions based on averaged mutual information performed better after several times of iterated segmentation. Specifically, differences between mutual information and averaged mutual information have been identified. While segmentation based on mutual information reached generally higher precision, non-collocations extracted with mutual information had generally larger edit distances than those extracted with averaged mutual information.

1 Introduction

Identifying collocations is an important part of preprocessing for multiple natural language processing applications, including word sense disambiguation, machine translation, and information retrieval, etc. (Lin et al., 2008). Such a task is especially important and challenging in Mandarin due to the lack of obvious word boundaries in Chinese orthography and its inherent nature of being highly compositional. While many segmentation tools, such as *jieba* (Sun, 2012) and *ckip* (Ma and Chen, 2003), can identify small-unit common collocations, their performances are largely affected when faced with domain-specific documents. Domain-specific larger-unit collocations often fail to be identified, resulting in less-than-ideal performances for subsequent tasks. This study therefore seeks to examine collocation extraction methods suitable for domain-specific texts in Mandarin.

While several past studies have proposed different collocation extraction methods in Mandarin (e.g., Xu and Lu, 2013; Li, 2007; Hui and Donghong 2008; Qian 2012), these methods all required the additional involvement of dictionaries or part-of-speech tags. While such methods are viable when dealing with common texts, a domain-specific dictionary is often unobtainable, and part-of-speech tagging also often fails when faced with domain-specific texts. As such, a purely association-rule-based method would be a more feasible solution for automatic domain-specific collocation extraction in Mandarin.

In this paper, we propose a novel technique for automated collocation extraction aimed at domain-specific texts. Specifically, we combine and compare two association measures, i.e., mutual information and its variant, averaged mutual information, with iterated segmentation, in an attempt to account for the changes in the frequency distribution at different levels of segmentation.

2 Methods

2.1 Corpora

In this study, a corpus consisting of 100,000 legal judgments (LC) ruled by Taiwanese courts was used. The documents were first preprocessed and then segmented into words with *ckip*.

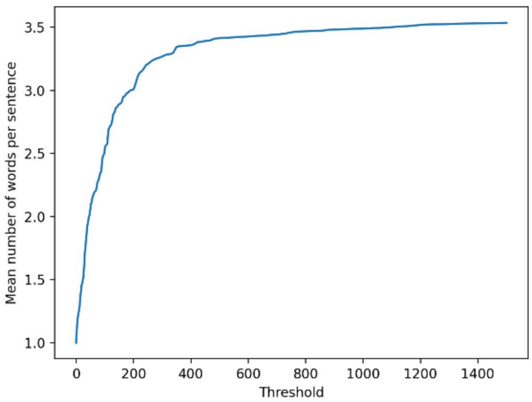

Figure 1: An example of the distribution of the mean numbers of words per sentence across different levels of MI threshold. The mean number of words stopped increasing at an MI threshold near 1500.

2.2 Iterated Segmentation Based on Mutual Information and Averaged Mutual Information

To perform iterated segmentation, the mutual information (MI) and averaged mutual information (AMI) of each pair of bigrams were calculated as in (1) and (2), where P(X) and P(Y) are the probabilities of X and Y, E[MI(X,Y)] stands for the expectation of the mutual information of X and Y, and H(X) and H(Y) stand for the entropies of X and Y.

$$MI(X,Y) = \sum x, y \in X, Y \; P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \qquad (1)$$

$$AMI(X,Y) = \frac{MI(X,Y) - E[MI(X,Y)]}{\frac{1}{2}[H(X)+H(Y)] - E[MI(X,Y)]} \qquad (2)$$

In each iteration, word boundaries were determined at bigrams with an (A)MI value lower than the segmentation threshold. To determine the segmentation threshold, the averaged numbers of words per sentence at different thresholds were calculated starting from 0 to when the averaged numbers of words stopped increasing (i.e., no words were segmented into a larger unit), with a step of 1 for MI and 0.001 for AMI. An illustration is shown in Figure 1. The elbow method was then used to determine the optimal segmentation threshold. The segmented words then underwent a new round of iteration, where the (A)MI values were recalculated. There was a total of 10 iterations.

2.3 Evaluation

To compare the interaction of different association measures and iterated segmentation, the extracted collocations after each iteration were evaluated (named MI-iterated 1–10 and AMI-iterated 1–10). To compare them with segmentation without iteration, 10 sets of collocations were additionally extracted without iteration, with segmentation thresh- olds based on the mean numbers of words at each level of the 10 iterations (named MI-noniterated 1–10 and AMI-noniterated 1–10). That is, the mean numbers of words of the noniterated groups were the same as their iterated counterparts. For instance, if the mean number of words for MI-iterated 5 was 3.5, then the collocations MI-noniterated 5 would be extracted based on the segmentation threshold at which the mean number of words for MI-noniterated was also 3.5. This was done to ensure the comparability of the iterated groups and their counterparts at each iteration level by making sure that they had the same mean numbers of words.

The extracted collocation candidates were then manually examined by four legal professionals. The candidates were labeled as three types: 1) legal collocations, 2) general collocations, and 3) non-collocations. Following Bouma (2009), label ranking average precision (LRAP) scores were used to evaluate the precisions of the extracted collocation candidates. Additionally, for candidates judged as non-collocations, the correct target collocations were also labeled by the examiners. Levenshtein distances were calculated between the non-collocations and the target collocations to estimate their similarities.

3 Results

3.1 Label Ranking Average Precision Scores

The LRAP scores of legal collocations and general collocations extracted with (A)MI- iterated and (A)MI-noniterated 1–10 are shown in Figure 2. As can be seen, an interaction between the different association measures and iterated segmentation was present. Specifically, for both legal collocations and general collocations, the extractions based on MI performed worse as the level of iterations increased; on the flip side, the precisions of the extractions based on AMI increased with the level of iteration.

The non-iterated groups, on the other hand, were less affected by the level of iteration. Specifically, the MI groups did not seem to be affected by the level of iteration, with the precision scores staying at 0.63 to 0.64 for legal collocations and 0.74 to 0.75 for general collocations

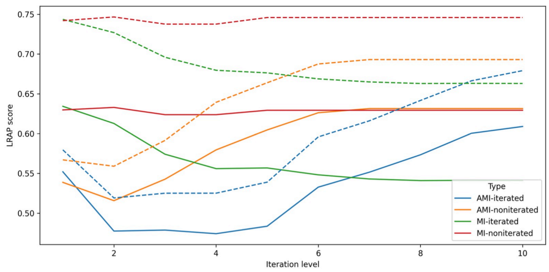

Figure 2: Label ranking average precision scores of the legal collocations (solid line) and general collocations (dashed line).

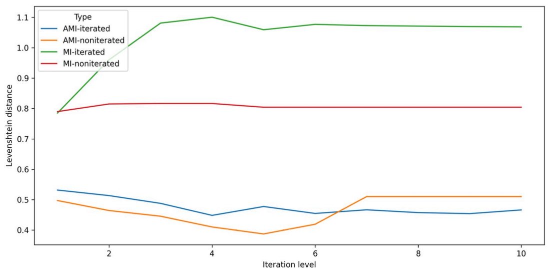

Figure 3: Levenshtein distance between non-collocations and target collocations.

throughout. On the other hand, the AMI groups performed bet- ter as the level of iteration increased. The precisions, however, stopped increasing after it reached 0.63 to 0.64 for legal collocations and 0.68 to 0.69 for general collocations as well.

Overall, the groups with the highest precisions were MI-iterated 1, MI-noniterated 1–10, and AMI-noniterated 7–10. For all groups, general collocation extractions had higher precision scores than legal collocation extractions.

3.2 Levenshtein Distance between Non-collocations and Target Collocations

The Levenshtein distances between non-collocations and target collocations for different extractions are shown in Figure 3. An obvious difference between MI- and AMI-based Domain-specific Collocation Extraction in Mandarin 5 extractions can be observed. In general, AMI-based extractions exhibited less edit distance between the false collocations and the target collocations than MI-based extractions. More importantly, while MI-based extractions, once again, did not benefit from iteration, the edit distances of AMI-based extractions decreased as the iteration level increased at the earlier stages of the iteration.

4 Discussion and Conclusions

4.1 The interaction between the precisions of association measures and the level of iteration

In Section 3.1, it has been found that MI and AMI-based extractions reacted to iteration differently. MI-based extractions did not benefit from iteration, while AMI-based ex- tractions increased in precision as the level of iteration increased. This might be due to the nature of AMI and its difference with MI. While MI measures the probability of two events happening together, AMI additionally takes into consideration the probabilities of one and both of the two events not happening. AMI therefore takes into account not merely the probability of the occurrence of a certain bigram, but also the counter- factual dependence of the two elements in the bigram, where the absence/presence of one element may promote the presence/absence of the other element. AMI-based ex- tractions may therefore be more sensitive to the change in probabilities of the co-occurrence as well as the counterfactual dependence of the elements in a given bigram after each iteration than MI-based extractions, while MI-based extractions may erroneously combine bigrams into collocations after several iterations without taking into account such counterfactual dependence.

4.2 Comparison of the performances between MI- and AMI-based extractions

Another issue worth discussing is the performances of the MI- and AMI-based extractions. The precision scores of MI-based extractions are higher than AMI-based ones at the previous stages of the iteration, and are close to AMI-based ones at higher iteration levels. As such, judging from precision scores, an extraction based on MI without ex- traction seems to be both more efficient and better performing. However, the edit distances of the AMI-based extractions were lower than MI-based extractions. Specifically, the edit distances of AMI-based extractions decreased as the iteration level increased at the earlier stages of the iteration. This suggests that AMI-based extractions may be a better choice if the purpose is to not only reach higher precision but also reduce the edit distances with the target collocations.

4.3 Performance ceiling of purely association-measure-based extractions

Another issue that is worthy of discussion is the performance ceiling of the extraction methods investigated in this study. In Fig. 2, it can be observed that whether it be iterated or non-iterated and MI- or AMI-based, the precisions scores for the legal col- locations seemed to stop increasing at a certain level (0.64). This might suggest that there exists an inherent limit to the performance of purely association-measure-based extractions. Indeed, in past studies, most collocation extraction methods require a combination of association measures and the additional involvement of dictionaries or part-of-speech tags. Extractions with high precisions may therefore be less attainable with purely association-measure-based methods. Alternatively, such a limitation may also surface from the relatively smaller sizes of the corpora used in this study, and the potential word segmentation errors during the initial segmentation process of the corpora. A larger corpus may disperse this question.

4.4 Conclusion

In this study, a novel association rule, i.e., averaged mutual information, and the use of iterated segmentation have been explored for domain-specific collocation extraction in Mandarin. It has been shown that compared with extractions based on canonical mutual information, those based on averaged mutual information benefited from iterated segmentation, though there seems to be a performance ceiling. Specifically, averaged mutual information has been found to reduce the edit distances between non-collocations and target collocations. The authors hope to provide further insights into the use of association rules in information retrieval, and to shed light on the issue of domain-specific collocation extraction.

5 Limitations

In the current study, only legal judgment texts are examined. It requires further investigation to determine whether the characteristics found in this study is applicable to other domain-specific texts.

References

Lin, J.-F., Li, S., Cai, Y. 2008. A new collocation extraction method combining multiple association measures. In *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*. IEEE, pages 12–17.

Sun, J. 2012. *Chinese word segmentation tool*.

Ma, W.-Y. and Chen, K.-J. 2003. Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics, pages 168–171.

Xu, R. and Lu, Q. 2013. A Multi-stage Chinese Collocation Extraction System. In *Advances in Machine Learning and Cybernetics*. ICMLC, pages 740–749.

Li, C. 2007. *Chinese collocation extraction and its application in natural language processing*. Ph.D. Thesis, Hong Kong Polytechnic University, Hongkong.

Hui, W., Donghong, J. 2008. Corpus-based extraction of collocations in Chinese. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, pages 330–333.

Qian, X. 2012. Automatic Extraction of Chinese V-N Collocations. In *Chinese Lexical Semantics*. CLSW, pages 230–241.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*. GSCL, pages 31–40.