Table 9: Results of our runs for Phase-3

System ID	TERM	SENTENCE	TF	COMBO	UNIQUE	MAP	Accuracy
KSU-JA-01@PH3	2/2	5/20	3/6	2/4	2/3	0/1	0.39(14/36)
KSU-JA-02@PH3	2/2	2/20	3/6	2/4	2/3	0/1	0.31(11/36)
KSU-JA-03@PH3	2/2	5/20	3/6	2/4	2/3	0/1	0.39(14/36)

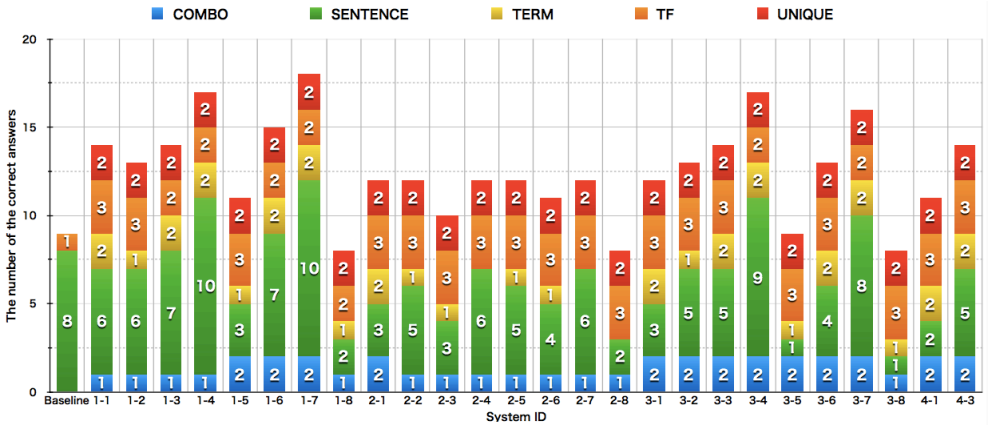

Figure 9: Comparison of the number of correct answers between the systems with different configurations

Figure 9 shows that the system 1-7 achieved the highest score of 18 correct answers, i.e. the accuracy of 50%, which used the OR query method and TWS. It also indicates that the system 1-4 and 3-4 accomplished the second highest scores of 17 correct answers, i.e. 47.2%, which used WS, and that the system 3-7 the third highest scores of 16 correct answers, i.e. 44.4%, which used TWS. Meanwhile, the score of Baseline was 9 correct answers, i.e. the accuracy of 25.0%, suggesting that the system 1-7 achieved improvements by 25 points. It was confirmed that the Baseline often gave wrong answers for the sub-questions with the answer type of TF and UNIQUE, whereas the system 1-7 chose correct answers for them. Thus, it seems reasonable to conclude that the query generation method in accordance to the answer types and other modifications successfully improved the accuracy.

For those systems using the OR query method and the VS query method, namely, the systems 1-1∼1-8 and 3-1∼3-8, respectively, the smaller the granularity of documents from Wikipedia became, the higher the accuracy of the systems got. This happened because 55.6% of all the sub-questions had the answer type of SENTENCE, and because it was easier to agree with the queries when searching for the sentence-based documents than for the paragraph- or page-based documents. Also, the sub-questions having the answer type of SENTENCE in the National Center Test almost always ask whether each answer choice is true or false, and these choices often have a brief description about historical events. Therefore, the sentence-based documents seem to have more documents which are more similar to those descriptions used in the answer choices.

For those systems using the AND query method, i.e. the systems 2-1∼2-8, the difference of the accuracy between the knowledge sources are smaller than those for systems using the OR or VS query. It is considered that the AND query method always returned the documents containing particular words, meaning that the scores of such documents became relatively high regardless of the granularity of documents. Thus, the systems using the AND query seem to have more often selected the same answer as a result. It was confirmed that the systems using the AND query actually selected the same answer more often than those using the OR or VS query.

6. CONCLUSION

This paper described the systems and results of the team KSU for QA Lab-2 task in NTCIR-12. In each phase of the task, we developed three automatic answering systems for world history questions in the National Center Test for University Admissions. In order for QA systems using document retrieval to answer questions correctly, it is important to estimate exact question types, and to utilize knowledge sources and query generation methods in accordance with these types. Therefore, we designed systems that focus on knowledge sources and query generations using the underlined texts in given exams. Scores of the formal runs were 20 correct answers(49%) and 68 points with KSU-JA-02@PH1 system in phase-1, 26 correct answers(41%) and 70 points with KSU-JA-01@PH2 system in phase-2 and 14 correct answers(39%) and 38 points with KSU-JA-01@PH3 system in phase-3.

7. ACKNOWLEDGEMENTS

A part of this work was supported by Kyoto Sangyo University Research Grants.

8. REFERENCES

[1] S. Hideyuki, S. Kotaro, I. Madoka, F. Akira, K. Yoshionobu, M. Teruko, M. Tatsunori, and K. Noriko. Task overview for ntcir-12 qa lab-2 (draft). In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 2016.

[2] H. Shibuki, K. Sakamoto, Y. Kano, T. Mitamura, M. Ishioroshi, K. Y. Itakura, D. Wang, T. Mori, and N. Kando. Overview of the ntcir-11 qa-lab task. In *NTCIR*, 2014.