Report on the Automatic Extraction of Korean Scientific Phrasal Term Candidates, with a Focus on Science Textbook Corpus

Jun Choi, Hae-Yun Jung, Hyeonah Kang and Seiyeon Kim
(Kyungpook National University, Korea)

In 2011, the International Association for the Evaluation of Educational Achievement (IEA) conducted the fifth 'Trends in International Mathematics and Science Study' (TIMSS 2011), in which fourth-grade students from 50 countries and eighth-grade students from 42 countries participated. The results of TIMSS 2011 revealed an interesting fact about Korean students. While they did achieve high ranks in scientific subject assessments, the rates of their interest in this field showed nonetheless the lowest scores.

Fang (2006) and Wellington & Osborne (2001) have suggested that the difficulty of scientific terminology was one of the major reasons why students find sciences hard and tend to avoid scientific studies. It seems in fact that scientific terminology constitutes a determining factor in students' interest in and understanding of scientific subjects. In that sense, we have been carrying out a project titled 'Development project of an integrated search system of educational science terminology through the construction of science terminology database'. In this paper, we present the procedure for automatically extracting scientific phrasal term candidates, which is part of the science textbook corpus building and term annotation processes, and report on the results of the science specialists and science education experts' examination of the list of candidates.

The list of morphosyntactic patterns (POS-gram) of scientific phrasal terms is compiled in two phases. A first list is obtained by analyzing the patterns of the 220,000 phrasal terms that are included in existing language and terminology dictionaries. Secondly, the list is refined by examining and identifying the conventional usage in the science textbook corpus during the morphological annotation process. This allows us to extract 437 types of POS-gram. This science textbook corpus consists of 6,000,000 words and is based on 132 science textbooks for grade 1 to grade 12 (i.e., for primary, middle, and high schools), used from 1992 to date.

The final list goes through a process of elimination so as to compile a list of stopwords, on the basis of which the POS-gram computation generates a list of automatically extracted candidates, which is then handed on to science specialists and science education experts for terminological annotation. Whenever a form of that list of candidates matches the one given by the dictionary, the respective specialised field information is assigned to it; the other terms are differentiated without any particular information being assigned. Finally, we discuss the issues related to the various scientific experts' examination and assess to what extent the automatic extraction of scientific phrasal term candidates is accurate and effective. As we investigated the feasibility of the automatic annotation of terms based on the morphological patterns in a given field, we could redefine the characteristics of that field and verify the emergence of new terms, thereby contributing to improving communication within the field.

References

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28(5)*, 491-520.

Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. Buckingham/Philadelphia: Open University Press.