AIOC System

Yu Gyung Kang
Deputy Director, Korea National Statistical Office, Statistical Information and Data Management
Dusan-dong, Seo-gu
Daejeon, Korea
summaul@nso.go.kr

1. Introduction

Korean standard industrial/occupational classification has been the basis of producing accurate statistical data related to our industrial structure and distribution of industry and occupation since 1960. But coding over several million records not only requires high cost in the aspects of time and manpower but also has much problems in accuracy and consistency. Therefore, we got to develop the automatic coding system in order to work out these problems of manual coding. We show the characteristics of our AIOC system and the result of experiment over survey data of 2,000 Census.

2. AIOC System

The AIOC(Automatic Industry/Occupation Coding) project launched in 1999 and completed in 2000. Its architecture is shown in Fig1.

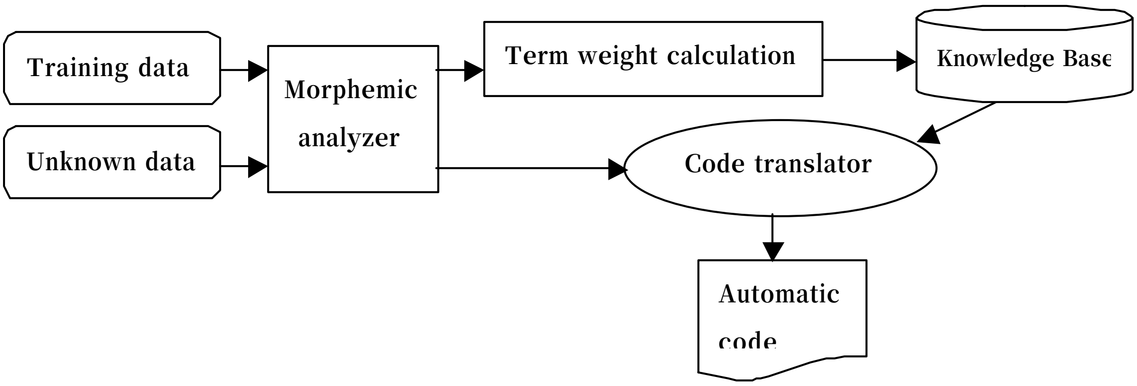

Figure 1. AIOC system architecture

The building process is mainly composed of two parts, i.e, the process of establishing a knowledge base and the process of making a code translator. We first have to establish a knowledge base and then we can classify unknown data by a code translator using the established knowledge base.

The performance of an automatic coding system is largely influenced by the knowledge base. As you can see in Fig 1, the training data flow to a knowledge base through a morphemic analyzer and a term weight calculation module, so the quality of a knowledge base is controlled by three factors, which are the quality of training data, the performance of morphemic analyzer, and the method in term weight calculation module.

The training data used for the knowledge base in the first version of AIOC system are the contents of the book named Korean Standard Industrial/Occupational Classification, about 60,000 records of past Census data and so on.

But this system has some drawbacks.

- It can't assign the relative importance to the words in the query

  For example, there is one record which says "making cars" in the main activity of company field. We know the word "making" is more important than the word "cars" to decide the industrial classification, but this system can't consider that.

- The method used in the term weight calculation , i.e tf(term frequency)×idf,(inverse document frequency) methodology, is not appropriate for this problem domain.

  The tf×idf algorithm is generally used for calculating the term weight. In this algorithm, the more frequently a certain word appears in a document and the less the number of documents where a word appears is, the larger term weight a word has. This characteristic is not appropriate for our problem domain The words such as "making" or "sale" appear in each 174, 54 documents in total document library of which size is 442. If we use the tf×idf method, those words have very low term weight in spite of their importance for classification.

- susceptibility of morphemic analyzer to spacing words

  Our morphemic analyzer is very sensitive to spacing words. It has the tendency of parsing sentences wrongly which are written without spacing words. The problem is that many of the unknown data we have are written without spacing words.

3. Experiment

We applied this system to the survey data of 2000 Census. We performed both manual and automatic coding, which could have us get the more accurate data and give the chance of evaluating the performance of our system. We assumed the records having same manual and automatic code were accurate, although there was some possibility they were misclassified. We reexamined and reclassified the mismatched records, which could improve the accuracy of coding.

The criterions of evaluating the performance of an automatic coding system are a production rate and an accuracy. The former means percentage of records that are automatically coded, and the latter means percentage of coded records that are well coded over total coded records. Unfortunately, we can't calculate the accuracy in strictly speaking because we can't know the exact codes over census records. We can just calculate the accuracy roughly assuming that all matched records are correctly classified and all reexamined/reclassified records are correctly classified.

The experiment over 1,939,257 records shows the production rate of both(industrial code, occupational code) is 99.8% and the matched rate of industrial code is 56%, and the matched rate of occupational code is 42%. The reexamination process is under the way and we can calculate the rough accuracy after that process.

4. Afterwards

The first version of our system has much to be improved and there is still much risk to apply it alone to other surveys. To improve the performance, we are building the second version of AIOC system to make up for the drawbacks of first version. Afterwards, we will continue researching new methods of automated coding.