Journal of Natural Language Processing Vol. 2 No. 2

A New Direction for Sublanguage N.L.P.

Satoshi Sekine†

There have been a number of theoretical studies devoted to the notion of sublanguage. Furthermore, there are some successful natural language processing systems which have explicitly or implicitly utilized sublanguage restrictions. However, two big problems are still unsolved to utilize the sublanguage notion: 1) automatic definition and dynamic identification of a text to sublanguage, and 2) automatic linguistic knowledge acquisition for sublanguage. There are now new opportunities to address these problems owing to the appearance of large machine-readable corpora. Although there have been several experiments to try to solve the second problem listed above, the first problem has not received so much attention. In the previous sublanguage N.L.P. systems, the domain the system is dealing with was defined by a human. This is actually one method to define the sublanguage of a text, and, in a sense, it seems to work well. However, it is not always possible and sometimes it may be wrong. In order to maximize the benefit of the sublanguage notion, we need automatic definition and dynamic sublanguage identification. We will report preliminary experiments on sublanguage definition and identification based on lexical appearance. The results of the experiments show that the methods proposed can be useful in processing a new text. In particular, the fact that the first two sentences can reliably identify a text's sublanguage encourages us in further investigation of this line of research. In conclusion, it appears that the inductive definition of sublanguage and sublanguage identification would be beneficial for natural language processing.

KeyWords: *Sublanguage N.L.P., Corpus-based N.L.P.*

† New York University, Computer Science Departent