Analyzing Mathematical Corpora to Improve Mathematical Handwriting Recognition

Stephen M. Watt

Ontario Research Centre for Computer Algebra
Department of Computer Science
University of Western Ontario
London Ontario, CANADA N6A 5B7
Stephen.Watt@uwo.ca

We are interested in the problem of mathematical handwriting recognition by computers. While some users may prefer a keyboard, many will find the pen to be a more intuitive way to express and manipulate mathematical formulae. Such an interface would be useful for many applications, from a math tutoring aid on an electronic white board to research mathematicians collaborating at a distance.

Disambiguation of character choices in the recognition of natural language text is usually aided by use of a dictionary. This is how "cloud" might be recognized instead of "cloucl" even when the first and last letters of the word look the same. In mathematics, however, there is in principle no fixed dictionary of multi-symbol words. Nevertheless, in practice, certain subexrpessions occur much more frequently than others. For example, the expression $\sin \omega t$ occurs frequently, while $\sin w t$ occurs almost never.

In earlier work [1] we have reported on the analysis of some 20,000 articles from the mathematics pre-print server arXiv.org. We were able to use the information gathered to construct *n*-grams to improve the recognition accuracy of a pen-based mathematics interface [2] producing MathML. The construction of *n*-grams from tree-structured data used a linearization technique to traverse the tree frontier and insert sufficient geometric symbols to keep track of the expression baseline.

In other work [3], we have made a similar analysis of the *n*-grams occurring in the domain of engineering mathematics, as taught to second year university students in North America. We believe this domain serves as a useful starting point for the mathematics that is used in practice by a significant population. We are currently studying how well *n*-grams can improve mathematical handwriting recognition in this context. The present article reviews the current results in this area.

[1] *Determining Empirical Properties of Mathematical Expression Use*, Clare M. So and Stephen M. Watt, pp. 361-375, *Proc.* Fourth International Conference on Mathematical Knowledge Management, (MKM 2005), July 15-17 2005, Bremen Germany, Springer Verlag LNCS 3863.

[2] *Mining Empirical Data to Improve Pen-based Mathematical Character Recognition*, Elena Smirnova and Stephen M. Watt, Communicating Mathematics in the Digital Era, (CMDE 2006), August 15-18 2006, Aveiro, Portugal.

[3] *An Empirical Measure on the Set of Symbols Occurring in Engineering Mathematics Texts*, Stephen M. Watt (submitted).