In this thesis, we present a data-driven system for disambiguating token and sentence boundaries. The implemented system is highly configurable and versatile to the point its tokenization abilities allow to segment unbroken Chinese text. The tokenizer relies on maximum entropy classifiers and requires a sample of tokenized and segmented text as training data. The program is accompanied by a tool for reporting the performance of the tokenization which helps to rapidly develop and tune the tokenization process. The system was built with multi-platform libraries only and with emphasis on speed and correctness. After a necessary survey of other tools for text tokenization and segmentation and a short introduction to maximum entropy modelling, a large part of the thesis focuses on the particular implementation we developed and its evaluation.